ABSTRACT

Title of Dissertation: COMPREHENSION OF CONVERSTATIONAL IMPLICATURE: EXAMINING EVIDENCE OF ITS SEPARABILITY AS A LISTENING SUBSKILL

Stephen P. O'Connell, Doctor of Philosophy, 2019

Dissertation directed by: Dr. Steven J. Ross, Second Language Acquisition

Understanding the inferences that speakers rely on to communicate is a core part of listening comprehension and is, more broadly, an important aspect of communicative ability. As a result, theories of communicative language ability account for it, and language testers who try to gauge the proficiency of learners of a second language include it in their assessments. Within the field of language testing, much research has been conducted to better understand how different aspects of listening may contribute to difficulty for second-language learners. One area of investigation has been the notion of listening being separable into different subskills, such as listening for inferences as opposed to listening for specific explicit details or listening for main idea. However, there have been mixed results when attempting to determine the psychological reality of these subskills.

This study attempts to clarify this question via a listening comprehension instrument that was designed specifically to assess the comprehension of conversational implicature, or pragmatic inferencing, in contrast to non-implicature, or general

comprehension. This balanced instrument was administered to 255 language learners in two item formats, multiple choice and constructed response. In addition, participants were administered short-term memory and working memory measures. A variety of analyses, including item response theory (Rasch), logistic regression, and confirmatory factor analyses were used to try to attain evidence for 1) the existence of a separable listening for conversational implicature subskill and 2) the validity of assessing this subskill through a multiple-choice format.

The results from the analyses generally converged to indicate that while conversational implicature contributes to difficulty, it is not a separable subskill. However, the results did show that the multiple-choice item format is a defensible method for targeting this skill. This leads to the conclusion that expending effort on assessing comprehension of conversational implicature in general language proficiency tests may not be necessary unless the test-use context places particular emphasis on this ability. Although it is an integral aspect of listening, from an assessment standpoint, performance on general listening items will likely give test users the information they need to make predictions about comprehension of conversational implicature ability.

COMPREHENSION OF CONVERSTATIONAL IMPLICATURE: EXAMINING EVIDENCE OF ITS SEPARABILITY AS A LISTENING SUBSKILL

by

Stephen P. O'Connell

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment of the requirements for the degree of Doctor of Philosophy 2019

Advisory Committee:

Professor Steven J. Ross, Chair

Professor Michael Long

Dr. Martyn Clark

Dr. Jared Linck

Dean's Representative: Professor Donald Bolger

Acknowledgements

I have many people to thank for what they have contributed to my reaching the point of completing this dissertation. To start, my time working in the Testing and Certification Division of the English Language Institute (ELI) at the University of Michigan was instrumental in my undertaking this task. Working with numerous knowledgeable and experienced language testers at the ELI opened my eyes to how much I had to learn about the field. In particular, because this dissertation focused on listening, I would like to acknowledge how much I learned from Sarah Briggs about assessing this not-so-easy-to-assess skill.

During my time in College Park, Maryland, my second language acquisition cohort and peers provided encouragement and support when it was much needed. Among them, I would like to thank Ilina Kachisnke, Katya Solovyeva, Payman Vafaee, Pete Osthus, Eric Pelzl, Alia Lancaster, and Megan Masters for often making stressful times less stressful. Additionally, completing this dissertation and my doctoral degree would not have been possible without the guidance and support of the University of Maryland's second language acquisition program faculty: Professors Mike Long, Robert DeKeyser, Kira Gor, and Nan Jiang. There are also many, many people to thank from the Center for Advanced Study of Language (CASL), where I learned an immense amount about conducting research.

For the dissertation project itself, there are a number of people I need to thank. For contributing their expertise and time to assist in the development of my listening test: Mark Chapman, Roger Frantz, Robin Stevens, Fabiana MacMillan, and Rachele Stucker. I also need to thank Fabiana MacMillan and Rachele Stucker for rating hundreds of

responses to the constructed response items on the listening test. Rachele Stucker, along with David Erdody, also helped with the creation of the audio for the listening test (a not unimportant aspect of a listening test). I also owe gratitude to Alex Berumen and Elida Berumen for help with the Spanish translations of various materials. And for help with data collection, I owe much to Pilar Sotelo and Giovanny Andres Ferreira Hernandez for their tremendous support and hospitality while I was intruding on their schools' time and space to run my study. Finally, much appreciation to Gad Lim, who read and gave detailed feedback on a rough early draft of the dissertation.

I also need to thank my committee—Martyn Clark, Mike Long, Jared Linck, and D. J. Bolger—for their careful reading of my dissertation and extremely helpful and thoughtful feedback. But above all, I need to thank my advisor, Steven Ross, who agreed to take me on as a student despite my knowing next to nothing about inferential statistics or proper research methods. I've learned an immense amount and greatly appreciate your guidance.

To my nieces and nephews, Orla, Cormac, Eamon, and Katherine, many thanks for always being engaging and entertaining interlocuters (and perhaps now your befuddlement over your uncle having "homework" will cease). And last but not least: Go raibh míle maith agaibh as ucht gach rud, Mam agus Dad. It took longer than it should have, but that's where the Connaught resilience comes in handy.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Listening is the primary way that the vast majority of people receive linguistic input for the first several years of their communicative existences and is a mode through which most people continue to receive and process vast amounts of information for the duration of their lives. The complexities of the process of transforming aural input into messages with meaning become apparent very quickly once researchers attempt to explain it. Theoretical explanations need to account for a large number of factors, including, but not limited to phonological variation, word- and phrase-level meaning, sentence- and discourse-level meaning, the speed of input, the role of cognitive variables such as attention and working memory, and the role of context and background knowledge. There are also different types of listening and different reasons for listening to consider. In the fields of second-language learning and assessment, disentangling these various factors, in the attempt to better understand what makes listening challenging for second-language learners, has been the focus of much research. One approach to better understand and assess listening comprehension has been the notion of subskills: understanding global meaning, understanding details, and invariably, understanding inferences. It is this last theorized subskill, understanding inferencing, and understanding conversational implicature in particular, that is the focus of this study, for understanding implicit meaning is widely believed to be at the core of being able to use a language effectively (Buck, 2001; Rost, 2011; Taguchi & Roever, 2017). The models of language proficiency proposed and discussed within the applied linguistics and second-language proficiency testing fields—which have directly informed construct definitions for listening and reading comprehension tests—have specified an ability to understand implicit meaning. Two influential examples are Canale & Swain's (1981) inclusion of sociolinguistic

competence (i.e., sociocultural rules of use and rules of discourse) in their theoretical framework of communicative competence and Bachman's (1990) inclusion of illocutionary and sociolinguistic competence in his theoretical framework of communicative language ability.

It is therefore well recognized that implicature is part of listening, and communication more broadly, but beyond that, it is widely believed that the ability to consistently understand implicature, both as a reader and as a listener, helps distinguish higher proficiency second-language (L2) users of a language from those who are still developing their abilities. This belief is reflected in numerous frameworks or scales of proficiency. For example, the Common European Framework of Reference (CEFR) includes "Recognizing implicit meaning" and "differentiating finer shades of meaning" at the top two levels of its six-level global scale (Council of Europe, 2001, p. 24) and the ACTFL Guidelines discuss the ability to understand implicit meaning at the "Distinguished" level, which is the top level of an 11-level framework (American Council on the Teaching of Foreign Languages, 2012). However, the evidence that exists to support this notion of implicit understanding as an "advanced" skill is ambiguous, and in regard to the CEFR, numerous researchers have cited its lack of specification (Alderson, Figueras, Kuijer, Nold, Takala, & Tardieu, 2006, Fulcher 2004, Weir, 2005a) as one of its limitations—when we look more closely for information on what is meant by understanding implicit meaning (particularly in the context of listening), there is little there to guide test developers. This leaves open the questions of what exactly is meant by "implicit meaning," and for second-language learners, whether we can disentangle

implicit comprehension's reliance on real-world knowledge from its reliance on linguistic knowledge, and finally, on what basis it is linked to higher level performance.

But there is little dispute that the ability to comprehend implicit meaning is vital to being an effective language user, and as a result it has been included in many proficiency assessments for decades. In many widely used language proficiency assessments, testing comprehension of implicit meaning has most commonly been operationalized through multiple-choice items (TOEFL, TOEIC, MET, TEAP, etc.). Probing the validity of testing this skill in this manner has resulted in much research, primarily of the sort that investigates the evidence that implicature is a skill that can be isolated apart from other theorized subskills, such as the understanding of an explicit global meaning of a text or the concrete details of a text. The results of this research have tended to provide qualified support for the notion of separability of skills (Freedle & Kostin, 1996; Kostin, 2004), although it is far from conclusive (Eom, 2008; Rupp, Garcia, & Jamieson, 2001; Song, 2008; Wagner, 2004). Even when factor analyses of large test-taker data sets point to the possibility of an identifiable and separate subskill of inferencing ability, the picture is inevitably clouded by the potential influence of other passage and item factors. In addition, some researchers on pragmatics have expressed doubts about the ability to test pragmatic understanding through multiple-choice items without trialing them first as constructed response tasks in order to obtain insight into actual learner pragmatic failures (Kasper & Rose, 2002, p. 98).

To try to tie these questions together, it is helpful to think about them in the context of work on test validity and score validation. As Kane (2010) wrote, the questions of validity are straightforward—i.e., "To what degree is one testing what one claims to be

testing?"—but providing the answers are less so. In this case, the premise is straightforward: test developers believe understanding implicature—i.e., the intended meaning of utterances that is apart from the surface or explicit meaning of the utterance—is an important part of listening comprehension; therefore, test developers choose to test it. But we need more evidence from different sources to support test developer claims about assessing implicature, and to consider the possible rebuttals to those claims.

Therefore a study such as this one, designed to include a measure that aims specifically to assess comprehension of conversational implicature, contributes to validity arguments for assessments of listening comprehension by looking at performance of implicature items across item formats (i.e., multiple choice and constructed response), by looking at its separability from general listening within item formats, and by looking at the relative influence of working memory, a cognitive factor that is widely understood to play a role in language processing (Juffs & Harrington, 2011; Linck, Osthus, Koethe, & Bunting, 2014). The necessary theoretical background that will be discussed in the chapter 2 literature review includes theories of communicative competence, the concept of implicature, theories of first language (L1) listening and L2 listening, research on assessing L2 listening, and the concept of working memory and its relationship with second language acquisition. Chapter 3 discusses the current study's research questions and how Kane's validity argument can be applied to the study. Chapter 4 describes the method used, Chapter 5 the results (broken into subsections by analyses), and finally, Chapter 6 provides a discussion of the results, including limitations and implications.

Chapter 2. Literature Review

As outlined above, delving into L2 listening comprehension processes, and in particular the comprehension of conversational implicature, requires a survey of the literature in a number of areas, including communicative competence and implicature, listening comprehension (in L1, L2, and how it is assessed), and working memory. These are addressed in turn in the sections below.

2.1 Communicative competence

Communicative competence, also referred to as communicative language ability, is often theorized to be divisible into two major components: 1) language knowledge and 2) ability for using that language knowledge (McNamara, 1996). Language knowledge is defined as knowing rules of grammar, knowing the meaning of vocabulary items, knowing the sociolinguistic norms of a language, etc. And while it may seem counterintuitive, this is the potentially more observable component of communicative competence, as there are techniques for eliciting that knowledge: tests of vocabulary and grammatical structures for example. The second component, the *ability* for using language knowledge, is the less easily defined or measured concept. However, the importance of this component for second language learners was made explicit in the applied linguistics field in 1972 by Hymes. How to model it (and what to call it) has since been much debated, but it has remained an essential part of the field's most influential models of second language proficiency (Bachman, 1990; McNamara 1996; Canale and Swain, 1981; Widdowson, 1983).

The term "strategic competence" has been used by several researchers to capture the "ability for use" component that Hymes brought to the fore. However, it has been theorized by different researchers in slightly different ways. Canale and Swain (1981)

used the term strategic competence in a rather specific way, as a description of the abilities a person had to "compensate for breakdowns" (p. 33). When Bachman (1990) discussed the ability for using language knowledge in his model, he labelled it as "strategic competence," and makes a convincing argument that strategic competence is brought to bear in all language-use situations, although due to a variety of factors, it is clearly stressed more in certain situations than others. For example, when a second language speaker lacks the vocabulary for an object that is crucial to her immediate need—or when a speaker (in one's first or second language) must deal with an unexpected question during a job interview. Bachman's definition of strategic competence seems to best match the component of communicative competence alternately called "ability for use" by Hymes (1972) and "skill" by Canale (1983), but is a bit more evocative of the ability being described. It is for this reason, and because it is one of the more current definitions for the "ability for use" component of communicative language ability, this is the term used in this study.

But whether the component is called "ability for use," "skill," or "strategic competence," the core notion is essentially the same: speakers of a language have variable knowledge of the language, and they have variable ability to deploy that knowledge in situations where communication is required. And the difference between these two different constructs—knowledge of language and the ability to effectively deploy that knowledge in real time—is seen both through anecdotal experiences (e.g., a language learner who can read newspapers articles in the target language on familiar topics when given adequate time but is unable to quickly compose text messages on the same topic without errors of grammar, word choice, tone, etc.) and empirical evidence of

productive speaking and writing skills lagging behind receptive skill ability of listening and reading, whether the learners are adults (e.g., Bloomfield, Ross, Masters, Gynther, & O'Connell, 2014) or children (e.g., Genesse, Lindholm-Leary, Saunders, & Christian, D., 2005). The anecdotal and empirical evidence both support the notion that building up grammatical and lexical knowledge of a language is one thing and being able to use that knowledge is another.

Much of the discussion of strategic competence, however, has focused on the productive skills of writing and speaking. This is not unexpected, as these are the skills where strength of strategic competence or lack of strategic competence is most clearly observed (i.e., the failure to respond to a question or to comment adequately), and often more directly problematic (i.e., immediate misunderstandings possibly ensue). However, depending on the communication situation, listening (and increasingly so with reading, with expanding use of real-time text/instant messaging to communicate) is also a skill that should be considered as part of strategic competence. This requires believing strategic competence to be the ability to conduct successful interactions (reliant on receptive and productive ability), and not just the ability to successfully complete only one's own side of the interaction.

This is of course not a novel idea. Kasper (1984) discussed the link between comprehension and communicative competence, and many of the rating scales for traditional speaking tests that involve human interlocutors include a category that pertains to listening ability. For example, Michigan Language Assessment's ECPE speaking test rating scale (Michigan Language Assessment, 2014) includes under the Discourse and Interaction category a "Listening comprehension" descriptor that says that at the top level

of the scale the test-taker "understands linguistic, sociolinguistic, and pragmatic information in order to engage in extended, spontaneous interaction." If the test takers' listening comprehension ability is not sufficient for understanding the input, successful communication is not going to occur. So it is not controversial to say that listening comprehension is one piece of a language user's strategic competence, particularly when we are talking about conversational implicature. Conversational implicature will be discussed in detail below, but in brief, being able to understand conversational implicature, the "pragmatic information" referred to in the ECPE Speaking scale, means being able to flexibly and rapidly use word and grammar knowledge in conjunction with language-use situations to obtain intended meanings; i.e., obtain some meaning from input and some meaning from context. As will also be discussed below, research has shown that conversational implicature is so widespread and is so automatic that proficient users of a language are rarely aware of it; it is an ever-present short cut used to increase the efficiency of communication. In fact, some researchers (e.g., Sperber & Wilson, 2012) argue that it is more than simply a short cut from the normal paths of literal coding-decoding, asserting that it is the means of alternate paths of communication (e.g., humor)—and that using it and understanding it properly is essential to having an adequate degree of communicative competence. If these views are shared by language testers, then it would seem to be important that listening tests include the construct of understanding implicature in their test specifications in a clear, specific, and demonstrable manner.

2.1.1 Implicature

The term "implicature" was coined by Paul Grice in his seminal 1975 paper "Logic and conversation." According to Wilson and Sperber (2012), this work was an attempt by

Grice, a philosopher of language, to bridge the gap between the two prevailing schools of thought in the field at the time: ideal language philosophy and ordinary language philosophy. To briefly summarize the difference, ideal language philosophers asserted that the meaning of utterances is within the semantics of the words expressed, whereas ordinary language philosophers claimed that the meaning of utterances is to be found within the context of the utterance (i.e., pragmatics). This is a bit of a simplification, but this is the context in which Grice was working and which he refers to as a "dispute" (Grice, 1989, p. 22–24). As Wilson and Sperber (2012) put it, Grice was attempting to reduce the gap between the two groups by explaining how it was possible to draw "a sharp distinction between sentence meaning and speaker's meaning, and explaining how relatively simple and schematic linguistic meanings could be used in context to convey richer and fuzzier speaker's meanings" (p. 1).

In his 1975 paper Grice stated that communication is only possible when those involved in the communication are operating under the assumption that the other participants are creating utterances that are good-faith attempts to be relevant and meaningful; that is, communication only works if all are operating rationally and cooperatively. Based on this foundational premise, Grice laid out the Cooperative Principle, which he subdivided into his four well-known maxims: 1) Maxim of Quantity, 2) Maxim of Quality, 3) Maxim of Relation, and 4) Maxim of Manner.

By Maxim of Quantity, Grice means that speakers should not say more or less than is necessary. The second maxim, the Maxim of Quality, refers to the notion that speakers should only say that which they know to be true and avoid saying that which is false or for which they do not have evidence. The Maxim of Relation addresses the

notion that speakers' utterances should be relevant. If Person A asks Person B what the temperature in the room is, Person B should not respond by saying what time it is (with Person A and Person B being strangers to each other since shared knowledge could play a role in providing meaning even in an exchange such as this). The final maxim is the Maxim of Manner, which was broken into four sub-components: i) avoid obscurity, ii) avoid ambiguity, iii) avoid unnecessary prolixity, and iv) be orderly. There appears to be some overlap between the Maxim of Manner (e.g., avoid unnecessary prolixity) and the first three maxims, but leaving that aside, these four maxims provided a framework for explaining how cooperative communication might happen in a logical, orderly, and efficient manner. Furthermore, as Cruse (2000) noted, the assumption is that these principles of cooperative communication are not culture-bound (making them of great interest to researchers of second-language use), nor are they relegated to the world of conversation and communication, but apply to any co-operative activity (p. 357–358).

But upon laying out these principles of cooperative communication, Grice immediately acknowledges that there is often a failure to fulfill these maxims. His 1975 paper lists four types of violations: 1) purposeful violations, 2) a conscious decision to opt out of fulfilling the maxims, 3) a clash between two maxims (i.e., the Maxim of Quantity calls for being as informative as possible but the Maxim of Quality calls for not saying that for which one does not have evidence; cases may arise where one is asked a question and there is tension between the two) and 4) the flouting of fulfilling a maxim based on the assumption that the failure to fulfill the maxim will *not* impede communication; i.e., that the relevance of the utterance will still be readily apparent even though a maxim is being flouted. It is this fourth type of violation—because of its

prevalence in daily communication—that led Grice to the notion of conversational implicature to account for it. And it is this violation and the notion of conversational implicature that are of primary interest for the current study.

Grice defines conversational implicature as involving three pieces. First, if Person A's utterance *p* implicates *q*, it requires that even though the letter of the maxims of the Cooperative Principle is not being followed, there is an understanding that spirit of the Cooperative Principle is in effect (i.e., Person A's intent is to communicate rationally). Second, it involves the belief that grasping the implication of *q* is necessary for utterance *p* to be seen as consistent with the Cooperative Principle. And finally, it means that person A "thinks (and would expect the hearer to think that the speaker thinks) that it is within the competence of the hearer to work out, or grasp intuitively, that the supposition in (2) is required" (Grice, 1989, p. 31). This essentially means that conversational implicature involves instances of communication where a maxim is flouted but is done so because the speaker does not believe that this will hinder communication. There are a variety of ways in which a maxim can be flouted, and this has led to attempts to categorize types of conversational implicature. Attempts by Bouton (1999) and others will be discussed below, but one major distinction should be highlighted first: *conventional* implicature as differentiated from *conversational* implicature.

Conventional implicatures are implicatures that follow logically or causally from the semantics of the utterance. One sub-category is entailment. The utterance that "John assassinated the Mayor" implies that John is a killer. Although it is not stated that John is a killer, there is no context or reading of the sentence "John assassinated the Mayor" that does not create the implicature that John is a killer because the semantics of the word

"assassinate" entails killing. Another type of conventional implicature is seen with stock phrases. For example, if Person A asks "Are you going to the party?" and Person B answers "Is the Pope Catholic?", it is understood that Person B is going to the party even though he did not say, "Yes, I will go to the party." There is no context or reading of the question "Is the Pope Catholic?" that does not translate into an utterance with the semantic meaning of "yes." The basis of conversational implicature, on the other hand, is the role that context and shared knowledge play in contributing to meaning. The work of relevance theorists such as Blakemore (1992) and Wilson and Sperber (2012), which is discussed below, is helpful in delineating how a listener arrives at the implicated meaning of an utterance.

While Grice created the term "implicature" to describe the intended meaning of an utterance that is not explicitly said, he chose to use the rather commonplace term "say" to describe utterances where implicature was not involved. It is in the work of the relevance theorists that the term "explicature" is used to contrast with implicature. Relevance theory arose, per Cruse (2000), because some researchers studying Grice's maxims concluded that the only Maxim that *cannot* be broken is the third maxim, the maxim of relation (or relevance), and that therefore all discussion of conversational implicature should center on relevance (hence the name "relevance theory").

Relevance theory, as proposed by Sperber and Wilson (1995), asserts that Grice's differentiation of what is said from what is implied is only partially helpful because it is extremely rare that what is said does not carry assumptions that must be interpreted based on the context and shared knowledge. That is, the use of implicature is the norm rather

than the exception. To use one of Sperber and Wilson's examples, the exchange in (2.1) is so uncommon that it would potentially be marked if it were uttered:

(2.1) M: Do you want to join us for supper?
W: No, thanks. I've already eaten supper tonight.

Responses that rely on assumptions are more common to conversational English. For example:

(2.2) M: Do you want to join us for supper?
(2.2a) W: No, thanks. I've already eaten.

Or

(2.2b) W: No, thanks. I've already had supper.

The response from W in (2.2a) conveys the assumption that she does not need to go into detail about what she has eaten or when she ate in order to provide a relevant, comprehensible response. The response in (2.2b) only requires an assumption about when. So it is the view of relevance theorists that a starting point to understanding implicature should not be that it is something that occurs only when certain maxims are flouted, but that it is something that occurs with great frequency in interaction (and automatically and unconsciously) and therefore requires a more subtle distinction than separating what is "said" from what is "implicated." It is following these premises that Wilson and Sperber discuss the notion of "explicature" and how it relates to implicature and the idea that the explicit aspect of communication relies as much on inferencing as the implicit aspect (Wilson & Sperber, 2012, p. 4). In other words, their theory helps provide a way to manage the observation that "…probably in the majority of cases [of speech], the propositions or proposition which constitutes the explicature are not fully

encoded in the explicit linguistic form" and that as a result, "the information that is conveyed through the overt linguistic form needs to be supplemented by processes of completion and/or enrichment" (Cruse, 2000, p. 352).

To discuss the additional explanation (the "completion and/or enrichment") that relevance theory provides, and which relevance theorists assert is part of the comprehension process, consider the following example from Blakemore (1992, p. 58):

(2.3) A: Did you enjoy your holiday?
B: The beaches were crowded and the hotel was full of bugs.

As Blakemore explains the Gricean view, the implicature is that speaker B did not enjoy his holiday. This implicature is made based on shared contextual knowledge between the speakers (i.e., few people enjoy crowded beaches or insects in their hotel rooms) and that the maxim of relevance is being followed. However, as Blakemore explains, the jump from speaker B's utterance to the implicature "B did not enjoy his holiday" is glossing over a step that should be explicitly described. That is, because the implicature goes beyond the words of B's utterance, we have to specify what a listener does to obtain the meaning beyond the words. This is enrichment of the actual utterance—the explicature—that Wilson and Sperber discuss. So in the case of dialogue (2.3), Person A formulates the following meaning from B's utterance:

(2.3a) The beaches at the holiday resort that B went to were crowded with people and the hotel where he stayed was full of insects.
(2.3b) B did not enjoy his holiday.

The explicature that takes place in (2.3a) is dependent on the meaning of the words uttered by B in (2.3), in conjunction with shared context and prior knowledge. The assumption, or implicature of (2.3b) is then derived from the fleshed out explicature. In this case, both the strength of the explicature and the strength of the implicature is high. It

is very unlikely that a large majority of competent listeners would not arrive at the same explicature and implicature in (2.3a) and (2.3b) based on (2.3). However, there are cases where either the strength of the explicature or implicature, or both, is weaker. Consider the exchange in (2.4), adapted from Wilson and Sperber (2012, p. 39):

(2.4) A: Do you want to go see a movie tonight?
B: I'm tired.

Here the explicature that B does not like to go out or does not like to go see movies when B is tired and that B is not likely to be refreshed enough in time for viewing a movie on the specific evening in question is based on few cues or little input and requires more enrichment—either from shared context or prior knowledge. However, again, the explicature, and ensuing implicature (that B does not want to go see a move that night) is likely available to the vast majority of listeners. As an example of an exchange where the implicature is even weaker, consider (2.5) adapted from Wilson & Sperber (2012, p. 15):

(2.5) A: I heard you moved from Manhattan to Queens.
B: The rent's lower.

In this case, the explicature involves enriching B's statement to mean that the rent in Queens is cheaper than the rent in Manhattan. But the step to implicature in this dialogue is not as clear. If one (fairly) assumes that A's utterance contains the implicit questions of "Why did you move from Manhattan to Queens?" or "How do you like living in Queens in comparison to living in Manhattan?" the resulting implicature is that cheaper rent is a reason to move and that it was B's only reason for the move. Further, because no other information was given, the implicature may also convey that otherwise, B is not thrilled about living in Queens in comparison to Manhattan. The implicature that B moved to Queens because she wanted to pay less in rent is competing with equally

plausible implicatures such as B can no longer afford to live in Manhattan and that the benefits of living in Manhattan were not worth the increased cost in rent. According to Wilson and Sperber, "the greater the range of [plausible] alternatives, the weaker the implicatures." This is one aspect of assessing implicature—through multiple-choice item format in particular—that language testers must pay close attention to because it is potential source of explanation as to why some implicature items are more difficult than others.

Delving further into the distinction between implicature and explicature inevitably brings up the challenging questions of where semantics ends and where pragmatics begins and whether all utterances are signals themselves or whether all utterances are simply evidence to help create a signal. As interesting as these questions are, they are beyond the scope of this study and its focus on language testers trying to assess conversational implicature. But a general understanding of the topic is of extreme relevance, in that it provides guidance to language testers in their attempt to cover the full domain of listening comprehension situations and in their attempts to describe what it is that listening comprehension tests are actually testing, i.e., the ongoing question of dividing comprehension ability into subskills. Asserting the position of Wilson and Sperber's view of implicature over Grice's, or the so-called neo-Griceans, is not the goal in describing their work. However, Wilson and Sperber's contributions to understanding implicature do seem to extend the implicature work that Grice started and place it in the more contextualized realms within which language testers (and all applied linguists) operate. Of particular interest is consideration of the steps required for a listener to get from an utterance to its explicature and then to its implicature. Attempting to identify and

quantify these steps—by judging the degrees of strength of both the explicature and implicature—and examine how they contribute to difficulty for language learners is an issue which will be returned to in Chapter 6.

2.2 Listening

As discussed above, listening is an important part of communicative competence and the ability to listen for implicature is part of that competence. Taking a step back from the notion of communicative competence, it is important to discuss the general act of listening. Listening is a multidimensional and complex process, a description of which must account for encoding, decoding, construction of meaning and how speakers/listeners and context relate to each other (Bodie, Worthington, Imhof, & Cooper, 2008). A number of theories have been put forth to try to explain the process. Testing these theories is not the focus of this study, but instead, they will be discussed in order to situate the present study within a complementary framework. I.e., listening for implicature is not occurring in isolation from other types of listening, so describing what researchers believe is happening with the overall listening process may help us to understand what is happening with implicature as well.

The initial step in the listening process is the neurological/physiological act of hearing. Rost (2011, p. 12–13) summarizes this as a process following these steps:

1. Sound waves are funneled from the outer ear into the ear canal.
2. Once in the ear canal, the sound waves vibrate off the eardrum.
3. The vibrations from the eardrum enter the middle ear, which is where the cochlea is located.
4. The cochlea is filled with fluid; the vibrations effect a movement of the fluid which creates an electrical impulse.
5. The electrical impulse is carried by auditory nerves to the brain.
6. The brain decodes and interprets the impulse as language (or as non-language sound).

Step six is the step in the process that listening researchers who are interested in understanding communicative listening have grappled with, as this is where comprehension actually occurs, i.e., how a listener can create mental representations from what is heard.

2.2.1 A model of listening comprehension

Kintsch and van Dijk (1978) proposed a model of comprehension, applicable to both reading and listening comprehension, that has influenced researchers in the years since and whose basic structure is also seen in alternative models. Their model can be summarized in three overarching steps: 1) meaning elements are organized into a coherent whole (i.e., a "text base" is formed), 2) the text base is condensed to gist, and 3) new texts are generated (as a response or summary).

In the first step, meaning elements are extracted from a text and those meaning elements are organized into a coherent whole. These meaning elements are found in multiple elements of a text. If we consider a listening text, meaning elements are carried in tone and prosody, the semantic information of lexical items chosen, and syntax (although Kintsch and van Dijk focused on semantic decoding and cohesion in their 1978 paper rather than syntax, they are clear that it is one of the aspects of a text that contributes to the organization of a coherent whole). Additionally, the context and shared knowledge of the text creator and the person comprehending the text play a role in organizing information coherently. What Kintsch and van Dijk, and others (e.g., Edwards, 2011) assert is that there are multiple parallel processes occurring. That is, neither the top-down nor the bottom-up ends of the processing spectrum are seen as viable explanations. It may very well be the case that in certain situations with certain listeners or readers, a bottom-up approach is employed to a greater extent (e.g., low

proficiency users of a language; a language user encountering a text that they have no background knowledge about) than a top-down approach, or vice versa, but it is rarely the case where there is no differential interplay of the two. The fact that multiple decoding processes are being applied to multiple meaning elements of a text at the same time is what leads to differential retention of different aspects of a text. This is because tone was used emphatically with a certain part of a text, it may be recalled more easily than other parts, or because one part of a text was delivered more slowly and clearly than the surrounding text it is recalled more easily, or because repetition was used with a certain phrase (e.g., in an advertisement or a speech) it is recalled more than others.

The coherent whole that is mentally composed by a listener or reader is Kintsch and van Dijk's text base; this text base is not an unrelated list of propositions (or information), but is instead a gist picture of what is being said. It is building and changing in real-time as new elements of new propositions support or detract from the coherence of the picture that the listener is forming—and the predictions that the listener is making. Problems occur when there is a failure to find a relational overlap between the content of new propositions and prior propositions—this is the aspect of a model of listening that would appear to be very relevant to the processing of implicature and will be discussed further below. Additionally, cognitive factors (e.g., working memory) also have the potential to constrain how well the gist picture is developed. The process of comprehension of multiple propositions being condensed to gist is one reason why a group of proficient language users might listen and accurately comprehend the same speaker in terms of overall message, but if asked to produce verbatim accounts of the message, will undoubtedly vary, even if the message was only a couple of sentences. The

final part of Kintsch and van Dijk's model, generating new texts, refers to the ability to create an appropriate response or summary of what one has comprehended.

This summary does not fully capture the intricacies of Kintsch and van Dijk's work, but it serves the purpose of providing a model of how listening comprehension is believed to occur, and is particularly useful in terms of examining conversational implicature by providing the notion of relational overlap of propositions to examine. Kintsch and van Dijk's model is echoed in more recent work by listening researchers, such as Bodie et al. (2008), who label these three main steps of the listening process as listening presage, listening process, and listening product. Bodie et al. describe listening presage as entailing personal (background knowledge, memory span) and contextual (purpose, interactivity, etc.) factors that are the preconditions to the listening event. The middle step, listening process, refers to the mental processes which are occurring, and the final step, listening product, refers to listening outcomes. The Bodie et al. attempt to provide a unified framework of listening is also useful for the examination of second-language listening in regards to the first and final steps. The specifics of person factors and listening context can be helpful in describing why one population performs better than others or why certain listening tasks are more difficult than others. Additionally, the inclusion of "understanding" as a listening outcome fits well with the notion of assessing understanding of listening texts in second language proficiency contexts.

Kintsch and van Dijk's and Bodie et al.'s models were created to address the question of how first language listening comprehension occurs. They are of course also directly relevant to second language listeners, because there does not seem to be evidence that second language listening occurs in a different way. In fact, when we look at the

work of researchers who are primarily concerned with L2 comprehension (e.g., Field, 2008; 2013) we see compatible theorizing about how meaning is created. Field (2013) describes "message processing" as entailing a complex system of decoding input at phonemic, lexical, syntactic and propositional levels simultaneously while integrating what is decoded into an ongoing discourse in a hierarchal (not linear) way. That appears to map with Kintsch and van Dijk (Figure 1, on left): the processing of multiple meaning elements simultaneously (1978, p. 363) can be seen to be parallel to Field's description of "message processing" and their "text base generation" is akin to Field's "integration of discourse in a hierarchical way" (Figure 1, on right).

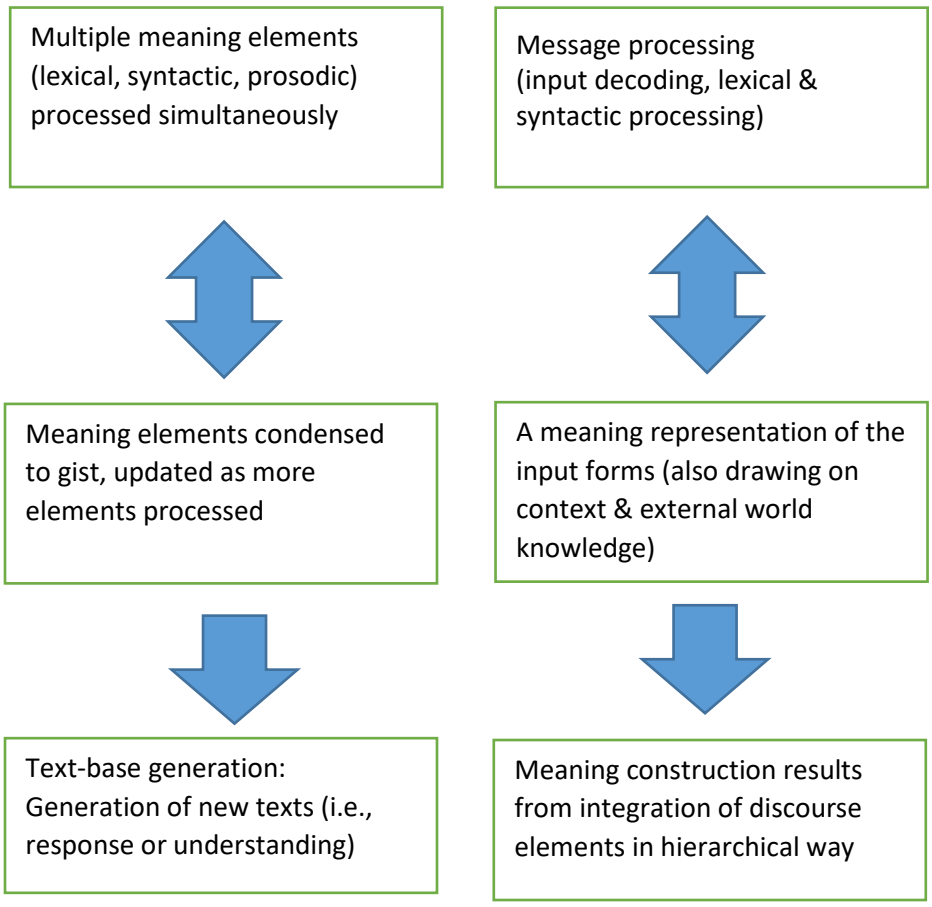


Figure 1. Summary of Kintsch and Van Dijk's and Field's Comprehension Processes

Controlled and automatic processes for decoding are necessary for both first-language and second-language listening, but for many second-language listeners, particularly those who are still developing proficiency, the extent of the lack of automaticity with what need to be automatic or nearly automatic processes is what limits them.

2.2.2 Listening assessment research

As stated above, many researchers have taken the view that first-language and second-language listening comprehension processes follow the same paths (e.g., Buck, 2001, p. 51; Clark, 2007, p. 40; Rost, 2011). But because becoming a proficient listener in a second language is a slower and more difficult process for many learners (particularly adults), it became clear to language teachers and applied linguists alike (Field, 2008) that assessments were required to provide estimates of this ability (which was not and is not the norm for L1 listening ability). The assessment of L2 listening comprehension lagged behind the assessment of reading comprehension by decades, however (Kasper, 1984, p. 2). As one example, per Field (2008), listening was not even included on Cambridge language exams—which have been in existence since 1913—until the 1970s (Weir, 2005b). And it was not until the 1980s that the listening texts on exams were actually spoken texts and not just written texts read aloud. The constraints of technology and greater logistical challenges of administering a uniform listening test certainly played a role; it was not until the advent of cassette tapes in the 1970s that it really became feasible. But in the ensuing decades, much research has been done on the question of assessing second language listening, with a great deal of it looking into the question of dividing the ability to comprehend a second language into subskills

This idea of different listening subskills (Buck, 2001; Dunkel, Henning, & Chaudron, 1993) informs many test developers' understanding of the listening construct and drives

the design of many second language listening assessments. It is also an assumption at the core of the present study—that a particular skill (understanding conversational implicature) can be targeted by design by test developers. Confirming that these subskills exist—by verifying hypothesized differences in difficulty of items that were designed to tap different subskills—has been the focus of much of the research.

Freedle and Kostin's (1996) study found an ordering of difficulty by subskill for the Test of English as a Foreign Language's (TOEFL) mini-talk listening items. They investigated four categories of listening item (main idea, supporting detail, inference, and application of inference) and found that the subskill categorization was a significant predictor of difficulty. Freedle and Kostin's follow-up study in 1999 (with the same dataset) sought to address the question of whether participants attend to listening passages while taking a multiple-choice listening test; that is, if participants are able to successfully respond to items without the benefit of hearing the passage, one of the criticisms of multiple-choice testing (a criticism, which if verified, is a serious threat to the validity of multiple-choice listening-passage item type). They looked at 337 items from 69 listening passages and categorized them by a number of different variables presumed to be related to difficulty. They then created a three-level progression of validity: difficulty stemming from item variables alone, described as a weak validity indicator; difficulty stemming from a combination of item and text variables, described as a medium validity indicator; and difficulty stemming from text variables alone, described as a strong validity indicator. Their conclusion states that their results provide evidence that refutes the criticism that items can be keyed without listening to the passages; that is,

the item variables appeared to play a smaller part in item difficulty than the item/text and text variables did.

Looking at another type of TOEFL listening item, short dialogues, Nissan, DeVincenzi, and Tang (1996) investigated 17 different variables to try to determine what aspects of the items and stimuli predicted difficulty. "Inference" was one of their variables (that is, they coded items according to whether the information being tested was explicitly or implicitly stated in the stimuli) and it was one of the five variables that were found to be significant indicators of difficulty. In another ETS TOEFL study, Kostin (2004) conducted a replication and expansion of Nissan et al. (1996). She included 49 variables in the study and the variable labeled as "an inference is required to respond correctly" was again found to be a significant predictor of difficulty. However, the main purpose of the TOEFL listening section is to assess one's proficiency as an "academic listener," (Sawaki & Nissan, 2009), so the primary focus of their research has been to establish the validity of the TOEFL as a predictor of success within that broader target language use domain, rather than to demonstrate that they are actually testing the ability to infer.

Turning to studies that investigated this question beyond the TOEFL, Eom (2008) used confirmatory factor analysis to test the hypothesis that through an analysis of 14 different variables, the listening comprehension section of the Michigan English Language Assessment Battery (MELAB) could be divided into two factors: language knowledge and comprehension, which would be in accord with theories of language proficiency (e.g., Bachman & Palmer 1996) that state that proficiency can be divided into language competence and strategic competence components. Eom's re-specified model

did seem to provide evidence for a cautious assertion of her hypothesis, that the multiple-choice items on the listening section of the MELAB do fall into either a language knowledge or language comprehension category, but the large number of variables and small numbers of items linked to each variable would suggest that more research would need to be done in this area. Additionally, it is hard to consider this study without wondering how all the items in the listening section are not in some way part of the comprehension factor (which is of course the challenge that all language testers face, identifying and operationalizing the line between knowledge and ability for use). Wagner (2004) also examined listening materials from University of Michigan language examinations, the MELAB and the Examination for the Certificate of Proficiency in English (ECPE). His results were also inconclusive. Only 13 listening items were included in Wagner's study and he concluded that since inferential understanding always entails a degree of explicit understanding, it will always be difficult to separate these as two distinct abilities.

Song (2008) used a structural equation modeling approach to look at both the listening and reading sections of a university English placement examination. Her primary question was whether the two receptive skills of listening and reading are a unitary trait, but in doing so, she examined the question of the existence of subskills in listening and found that a three subskill categorization (main idea or main topic of a text, understanding supporting details, and making inferences from explicitly stated information) was supported by her model. However, it should be noted that the listening assessment portion of her experiment consisted of only one aural stimulus and the 20 items based on that stimulus.

Rupp et al. (2001) set out to investigate the question of subskills by examining the sources of item difficulty, with the belief that understanding sources of difficulty in items will aid test developers' understanding of the listening construct. In their study, 87 non-native speakers of English (with a variety of L1 backgrounds) responded to 214 multiple-choice items. Although the main focus of their study was to compare the results of two different statistical methods, their multiple regression analyses found that passages with longer sentences, higher word counts and higher type-token ratios were more likely to be difficult. Most relevant to the study at hand, their analysis of items by type (main idea, detail, prediction, and understanding relations) were significant predictors of difficulty using multiple regression, with main idea items being the easiest (although their separate classification and regression tree (CART) analysis found it less significantly so than in the multiple regression). However, they ran the analyses of reading and listening items together and independent results for the two skills were not reported.

Summarizing these studies, it appears that on the whole there is some evidence for the divisibility of the listening skill. But what many of these studies on listening subskills lack is a precise definition of each subskill. Items that are classified as "inference items" are often a mixture of overall gist, speaker attitude, prediction, conventional implicature and conversational implicature. Part of the reason for this lack of precision is the result of research that has utilized existing test forms and test data. This also results in analyzing datasets with numbers of items by subskill that are not balanced (i.e., many explicit detail items, few inference items, and even fewer main idea items). Working with a precise definition of a subskill such as "pragmatic listening" as it is interpreted by item writers and then testing it in sufficient numbers to make generalizations about its difficulty is

what is missing from the assessment literature. A small number of language acquisition researchers—working primarily outside of assessment—have tried to understand conversational implicature more thoroughly; their work is discussed below.

2.2.3 Research on comprehending conversational implicature

Despite a growing literature in the areas of second language listening and assessing second language listening, as both Roever (2013) and Taguchi (2005) have noted, exploring the pragmatic aspects of listening has not received much attention despite the fact that incorporating pragmatic competence (or strategic competence as it has been alternately known) in models of communicative competence is widely accepted. Much more work has been done on the ability of learners to *produce* pragmatically appropriate responses to certain utterances or in certain situations. However, the results of those lines of inquiry, while important, have been conflicting, primarily because there is more variety in how people respond to certain situations than in how they interpret them. That is, if we asked 100 proficient users of English how they would decline an invitation to a party that they didn't want to attend, the range of responses would be far wider than the range of interpretations generated by asking those same 100 proficient language users when asked about the meaning of a response such as "Oh really? Saturday's pretty crazy for me" to the utterance "I'm having a barbecue this Saturday afternoon. You're welcome to come." So it could be argued that for learners of a language, determining the ability to comprehend the intended meaning of an utterance is an essential step. Work in this area was spearheaded by Bouton (1988). His 1988 study was among the first to look at the ability of non-native speakers of English to understand implicature. However, it must be emphasized that even though he was investigating conversational implicature, all of his materials were presented to participants in a written format. But putting that aside, the

first important finding from his 1988 study was that implicature can be assessed using multiple-choice items. That is, that test developers can create situations and dialogues with a constrained degree of context that lead to extremely high rates of agreement among native speakers on what the communicative intent of the speaker was. Bouton's test included 31 items. For 17 of the items, 93% or more of his 28 American English native speakers agreed on the inference. Another 6 items had 79–89% agreement. Five had only 64–75% agreement and only 3 items had 50% or less agreement. The 23 items that had 79–100% native-speaker agreement speaks to the aspect of Buck's (2001) default listening comprehension construct that entails making "whatever inferences are unambiguously implicated by the content of the passage" (p. 144). The 8 items that had weaker agreement, and the 3 that had less than 50% agreement should be items, in the context of high-stakes standardized test development, that are either revised and fixed by developers and re-piloted, or discarded because of the lack of a natural or automatic (i.e., "tip-of-tongue") response.

Additionally, Bouton went to pains to create a variety of implicature type items, which he linked back to Grice's work by taking into consideration which of the different maxims were being flouted. Table 1 contains a summary of how Bouton classified his implicature items.

Table 1. Bouton's (1988) Items by Implicature Type

Implicature Type	**Implicature sub-type**	**Number**
Relevance		13
POPE Q*	Relevance	2
Quality		1
Change subject	Relevance	1
Irony	Quality	4
Manner: Sequence		2
Quantity		4
Understated negative evaluation		4
TOTAL		**31**

*POPE Q refers to conventionalized lexical phrases that have one unchanging meaning, e.g., "Is the Pope Catholic" utterance in place of a "Yes" utterance.

His taxonomy and the study's results led him to conclude that certain types of implicature, i.e., understated negative evaluation, were more difficult than others for NNS. However, the lack of balance in his implicature types makes it difficult to make such an assertion with confidence. Additionally, pinpointing the differences or the meanings of the types is not always easy (for example, the author found the "manner: sequence" type items to be confusing or contrived). Finally, there clearly are different degrees of strength of implicature. That is, it is more likely that the number of cues that require explicature and implicature is what makes a particular utterance more or less difficult than a specific "type." Returning to Buck's (2001) default listening comprehension construct, what is the line between inferences that are unambiguously implicated by the speech and those that are "ambiguously" implicated? But from the perspective of assessing the ability to comprehend implicature, it is very unlikely that high stakes tests will include specifications that call for a distinction between implicature types in a way that is consistent and meaningful (but admittedly, it is also unlikely that developers of standardized tests will have the time or resources to consistently and reliably quantify the number of implicature cues either).

Bouton (1994a, 1994b, 1999) continued his work with comprehending conversational implicature with research that looked at the learnability of this skill. He concluded that increasing proficiency and exposure to the target language led to better performance, but specified a difference between conventional implicature (the Pope Qs) and conversational implicature. Conventional implicature, he asserted, being more dependent on lexical items or pragmalinguistic structures, was easy to teach but difficult to acquire, whereas conversational implicature, being more dependent on context, was difficult to teach but easier to acquire. He also concluded that exposure to the L2 was more important than proficiency for acquiring the ability to comprehend conversational implicature.

Roever (2006, 2013) is another researcher who has investigated assessing the ability to understand conversational implicature. His 2006 paper describes the development and validation of a web-based test of ESL pragmalinguistics. The three subcomponents that he assessed were comprehending routines, speech acts, and conversational implicature. Focusing on his conversational implicature enquiry, as they are most relevant to this study, his measure contained 12 items; these items were taken from Bouton (1999) and he subdivided the 12 items into only two categories: idiosyncratic implicature (i.e., conversational implicature) and conventional implicature. As with Bouton's original study, the conversational items were delivered in the written modality. The measure (taken by more than 200 learners) had good reliability (Cronbach's alpha = .82 and Kuder-Richardson 21 = .80) but correlated only moderately with the routines and speech acts measures (.32 and .48 respectively). Roever was able to explain this because success with routines, in particular, depend more on knowing a

certain type of exchange rather than being able to apply language knowledge to a particular dialogue to discern what is meant by a given idiosyncratic implicature. In fact, what he found with implicature is interesting: contrary to Bouton (1988), his findings showed that proficiency did in fact have a strong predictive effect on comprehending conversational implicature, more so than exposure (which was defined as time spent in the L2 environment with native speakers of the L2, rather than time spent studying the L2).

Roever (2013) reports the results of an additional study that included a conversational implicature component (k=10). The assessment in question was a communication skills test for students interested in pursuing a health science degree at universities in Australia. This meant that the test-takers included native speakers of English (n=223) and non-native speakers of English (n=212) whose proficiency was strong enough to have not impeded their acceptance at tertiary educational institutions. This fact, along with the fact that conversational implicature (as it was in Roever (2006) and the Bouton studies discussed above) was tested through written material, is what resulted in near ceiling performance for the native speakers and extremely high performance from the non-native speakers (7 of the 10 items were responded to with > 80% accuracy and 2 items with > 70% accuracy). The one item that the non-native speakers struggled with (44% accuracy) was also moderately difficult for native speakers (57% accuracy), which would indicate a potential problem with the item. In the discussion of that study, however, Roever's questions regarding the absence of norms and benchmarks in pragmatics testing should be of great interest to language test developers and researchers. As mentioned above, scales developed by the CEFR and the ILR assert

that the ability to comprehend conversational implicature is considered a higher-level skill. Roever's 2013 results from advanced non-native speakers do not dispute that, but the question of to what degree this skill is available to intermediate learners, is unclear and is likely to remain difficult to answer without studies that focus directly on comprehension of conversational implicature through aurally delivered materials.

Taguchi (2005, 2007, 2009) is one of the few researchers who has investigated comprehension of conversational implicature through aurally delivered material. Her 2005 study looked at the effect of different types of implicature on accuracy and speed of comprehension, the effect of proficiency on comprehension of different types of implicature, and the relationship between accuracy and comprehension speed. She found that for her 160 adult learners of English (in a foreign language setting, Japan), the type of implied meaning did have an effect. She divided her 32 implicature items into 16 "more conventional" and 16 "less conventional" items. It is important to note that by conventional, she does not solely mean lexicalized chunks, as Bouton did (e.g., "Is the Pope Catholic?"). For example, the dialogue in (2.6), from Taguchi (2005, p. 549), was classified as conventional implicature.

(2.6) Jane: Do we have time to go over my paper?
Dr White: Oh, ah, do you mind if we talk about it tomorrow?

In this dialogue the tested implied meaning is that Dr. White wants Jane to come back the following day. This is considered more conventional because the implied meaning of "No, I don't have time" is delivered through the conventionalized polite routine of offering an alternative, in combination with the conventional lexical chunk "do you mind" as a way to introduce something that might not be preferred. An example of a less conventional implicature item from Taguchi (2005, p. 549) is shown in sample (2.7):

(2.7) Dave: Do you like the people upstairs?
Susan: We're always visiting each other.

The implied meaning that is sought here is that Susan likes the people upstairs very much. This is a slightly unnatural dialogue, however, as a "Yes" or "Yeah" would likely be uttered before expounding on how often they visit each other (unless intonation is used to account for the absence of the expected "yes" or "no."). It is therefore not surprising that the more conventional items were easier than the less conventional items. But while experienced item writers might find fault with some of the items, there are important differences in the degree of strength of an implicature to be considered, which Taguchi is drawing attention to. That is, items (2.6) and (2.7) both demand implicature of the listener to get at the desired meaning, but the difference in degree is quite wide, and if at all possible, it is in test developers' interest to be able to discern weak implicature and strong implicature items in their development and analysis processes.

Taguchi continued investigating aural comprehension of conversational implicature with a 2009 study that looked into the potential of corpus-informed input to lessen the type of somewhat contrived exchanges of items like sample (2.7). However, even after searching through two corpora that contained only face-to-face spoken interactions (the Santa Barbara Corpus of Spoken American English and the TOEFL 2000 Spoken and Written Academic Language Corpus), the dialogues still had to be "adapted to better serve the level of the target learner group and goals of the research" (Taguchi, 2009, p. 742). Ultimately, she was unable to use truly authentic exchanges as those types of exchanges rely on so much shared context that they are almost impossible to base items on in a comprehension testing situation.

The items that Taguchi (2009) developed and used were first trialed as constructed-response items with 10 native speakers and the majority "showed relatively uniform interpretation" (p. 742). This is an important step in validating implicature items and it raises the question of how well non-native speakers might handle the comprehension task if it involves constructed response rather than selected response/multiple choice (Taguchi did not have the NNSs respond to the items as a constructed response task). The 2009 study ended up with 32 experimental implicature items, 16 classified as more conventional and 16 as less conventional; it appears that this distinction was based solely on the speech act being performed by the speaker: indirect refusals were considered more conventional and indirect opinions less conventional, rather than a consideration of the number of cues or amount of context provided in support of the intended meaning. Her 2009 results with the new items mirrored the 2005 results: native speakers performed near ceiling on all the items (both more conventional and less conventional implicature) and non-native speakers performed better with the more conventional implicature than the less conventional implicature items. In an extension of her earlier studies, the 2009 study contained a longitudinal dimension, and when non-native-speaking participants were tested a second time (after a duration of a college semester in Japan), they showed improvement with conventional implicature (indirect refusals) but less improvement with routines, which are also more conventional.

However, when discussing implicature, there is the issue of separating real-world knowledge from language knowledge. Consider the following example:

(2.8) A: I thought you were going to cut the grass today.
B: It rained last night.

In this exchange, the implicature is that B is not going to cut the grass "today." This implicature is attained by A by formulating the explicature that: It rained last night and that has resulted in the grass being wet; wet grass is difficult to cut—it sometimes causes problems for the lawn mower. Therefore, B will wait for the grass to be dry before cutting it. Attaining this implicature requires real world knowledge of the effect of wet grass on some lawn mowers. As a result, this is a type of exchange that could not be fairly included as a dialogue in a standardized listening test—the percentage of people who share the real-world knowledge about wet grass and its effect on lawn mowers is too small. Thus a question often raised about implicature is whether it is a language issue or a real world knowledge issue, and the short answer is that it involves both. If A does not know the meaning of the verb "rain," he will not obtain the implicature, and if A does not know about B's preference for not cutting wet grass, he will not obtain the intended implicature.

The view then, for those from a more purist pragmatics perspective, i.e., those who adhere to the notion that meaning only emerges in context, is that implicature cannot be assessed in standardized tests. But in terms of item development for standardized language tests, the assumption is that implicature items can be created that do not rely on specific real-world knowledge—but instead rely on a type of general real-world knowledge. From a practical perspective, meaning can often be gleaned by those from outside the context. This is one of the purposes of the review and piloting process: to flag an exchange like (2.8) for revision or removal from a test. This is the same issue that standardized test makers must address with all material selection (writing task prompts, reading passages, etc.): ensuring accessibility to a wide range of test-takers without

advantaging or disadvantaging sizeable numbers. But this then raises the question of whether the implicature items that *are* included, are really assessing implicature if the real-world knowledge is accessible to all. It is the author's view that they are testing implicature and that they are testing this ability without testing real world knowledge. They must, inherently test *some* real-world knowledge, but the assumption is that it is accessible real-world knowledge: all test-takers have encountered unpleasant people or have been too busy to attend a party, so criticisms of such people delivered via implicature or declining invitations via implicature should not be inaccessible.

But even assuming that test developers can test this aspect of listening, several aspects of investigation of this practice appears to be missing from the literature. One is exploring the issue of assessing conversational implicature within the context of standardized second language tests through a study that tries to look at degrees of implicature based on Kasper's (1984) description of general context, specific context, and number of provided indictors as a way to consider difficulty rather than by speech act category. Another gap that this study attempts to address is to provide validity evidence for a common method of assessing the ability to comprehend conversational implicature: multiple-choice items. In'nami & Koizumi's (2009) meta-analysis of the multiple-choice/constructed-response contrast showed that this difference in format has been the subject of multiple studies in the domains of L1 reading, L2 reading and L2 listening comprehension. However, it does not appear to have been the focus of any studies that looked specifically at implicature, and those that look at L2 listening more generally are few. Their analysis included only five studies that looked at the difference in format effects between multiple-choice and open-ended L2 listening comprehension questions,

and none of those included implicit vs. explicit questions as a variable (In'nami & Koizumi, p. 236). The topic therefore calls for further exploration, as multiple-choice conversational implicature items are both a common feature of L2 listening tests and are an integral part of the assumptions about what differentiates higher proficiency from intermediate proficiency learners; further evidence for this assumption and for the validity of assessing implicature with multiple choice is required. The following section discusses the inclusion of measures of working memory capacity as a way to possibly help explain whether the relationship between comprehending implicature is more tied to general language proficiency or if cognitive factors also play a prevalent role.

2.3 Working memory

While the role of cognitive factors in acquiring and using a second language has been the subject of a great deal of research in recent decades, investigating how cognitive factors, such as working memory, specifically contribute to the ability to understand conversational implicature in a second language has received far less attention.

But before summarizing some of the research on working memory and second-language learning, it is first necessary to discuss the mechanism, or concept, itself. The Baddeley and Hitch 1974 model of working memory is probably the most influential and widely accepted. Their initial conceptualization was of a three-component system that is capable of storing and processing in the mind input from the external world. The three components were a central executive and two input systems, the phonological loop and the visuo-spatial sketch pad, which send stimuli to the central executive, where cognitive processing is then performed. As Caplan, Waters, & Dede (2007) summarized it, the model was "a capacity-limited, short-duration store in which computations are performed in the service of task goals" (p. 272). Baddeley (2000) later updated the model with the

addition of a fourth component known as the episodic buffer. The episodic buffer was how working memory's interaction with long-term memory (LTM) was theoretically introduced. It was conceived as a "temporary store in which the various components of working memory…can interact through the participation in a multidimensional code, and can interface with information from perception and long-term memory" (Baddeley, 2010, p. R138).

Although likely the most widely known and cited, the Baddeley model is not the only model of working memory in the cognitive psychology literature. Another prominent model is Cowan's (1988) embedded processes working memory model. What distinguished the Cowan model from the Baddeley model, especially before Baddeley's addition of the episodic buffer, was its placing of working memory processing within LTM. The rationale was that "[a]t any moment there is assumed to be a currently active subset of long-term memory, and the focus of attention is assumed to be a subset of that activated information" (Cowan, 1999, p. 88). Therefore, working memory cannot be used without implicating long-term memory as well. This assumption seems especially appropriate in the context of L2 comprehension of conversational implicature. Once learners attend to L2 aural input, it is theorized that in their attempt to process it, they are relying on what they have learned, that is, what is represented or stored in their long-term memory, in order to create the explicatures that are necessary for attaining a correct implicature. Failure with an L2 listening implicature test task can occur because there is nothing in the LTM that links to the input temporarily stored in WM, i.e., the lexical item or grammatical structure is new. Lack of comprehension can also occur because the representations in the LTM are not strong enough to allow the processing to occur at the

speed that is necessary before the lexical items or grammatical structures being attended to drop out of working memory storage. Cowan's model proposes a more general approach than Baddeley's model in that it does not specify how different types of stimuli are fed into the system (in Baddeley's model, through the phonological and visual-spatial loops), and Cowan (2005) concedes that his model can be criticized for leaving these areas unspecified. But its inclusion of the concept of LTM "activation" and time limits versus capacity limits seems more relevant, or at the very least more intuitive, to a discussion of L2 language processing where phonetic, syntactic, semantic and pragmatic information from long-term memory must be activated while time is a serious constraint (especially in the context of speeded listening comprehension assessment tasks). Additionally, Baddeley (2007) writes that the search for a link between the central executive and long-term memory is one of the key questions of this area of research.

Although there are variations of the WM model in the literature (see Gathercole, 1996; Conway, Jarrold, Kane, Miyake, & Towse, 2007), most researchers agree with the core principles that WM comprises domain-specific storage and domain-general executive control (Williams, 2012). But among the various models of WM in the literature, for the purpose of this study, WMC is being conceptualized in terms of the Cowan model for the specification of the ongoing and automatic interaction with LTM that it entails.

As mentioned above, the role of WMC in language processing, has been the focus of much research and there is considerable evidence to support the belief that working memory capacity is a predictor of ability in a first language (Daneman & Carpenter, 1980, 1983; Ericsson & Kintsch, 1995; Daneman & Merickle, 1996), albeit some

findings have included caveats or restrictions. For example, its role seems to be more prominent with L1 vocabulary development in children than in adults (Gathercole & Pickering, 1999; Gathercole, Service, Hitch, Adams, & Martin, 1999) and its role has not been seen clearly with L1 syntactic processing (Caplan et al., 2007), leading to an interpretation that some L1 processes are so automatized that there is little burden on WMC. However, both of these aspects of the WMC and L1 research—the connections and the limitations—have made it an attractive area of investigation for L2 researchers. Regarding the former, there has been a desire to see if similar predictions between WMC and L2 ability also could be made. Regarding the latter, because L2 processing is different in manner but not in kind from L1 processing, with the majority of L2 learners not reaching full expert level, the fact that WMC is not involved as strongly in certain areas of L1 use and ability is a possible signal that it will be prominent in L2 use and ability (e.g., for a skill such as comprehending conversational implicature). Supporting this notion, some studies have indicated that degree of L2 proficiency does make a difference in the extent of WMC's impact (Cheung, 1996; O'Brien, Segalowitz, Collentine, & Freed, 2006). As a result, L2 researchers have been investigating the belief that greater WMC will allow more noticing of the linguistic features of input, and with the ability to notice, the more will be learned (Ellis, 2001; Sawyer & Ranta, 2001).

An early study of the role of working memory in L2 development was conducted by Harrington and Sawyer (1992), in which Japanese learners of English were given a series of memory tests and a modified version of the reading span test in English. Scores on participants' TOEFL grammar and reading sections did correlate with the reading span test, but the fact that the reading span test was in the learners' second language introduced

a possible confound. Regardless, the correlations between L2 ability and WMC provided sufficient grounds to continue investigating this connection, albeit while attempting to measure WMC with tasks that minimized potential overlap with either the existing L1 or L2 abilities of study participants.

Additionally, working memory, and specifically its phonological component, has been strongly linked to L1 lexical acquisition. To investigate whether the same link could be found with L2 learners, O'Brien et al. (2006) examined the potential role of phonological short-term memory (PSTM) in L2 lexical acquisition and L2 speech production. Their study was conducted with a group of English-speaking university-age learners of Spanish over the course of a semester. PSTM was measured with a serial non-word recognition task (using non-words to minimize the influence of existing L1 or L2 lexical knowledge). They found that PSTM did play a significant role in the learners' development of narrative abilities and correct use of subordinate clauses. However, when comparing the lower proficiency cohort with the higher proficiency cohort (determined by a proficiency exam at the start of the semester), the results were mixed. That is, PSTM did not account for differences in performance by the lower or higher ability groups. Mackey, Adams, Stafford, & Winke (2010) also looked at speech production and WMC capacity. They investigated whether WMC was a significant predictor of the ability of 42 English-speaking learners of Spanish to attend to and incorporate into their spoken Spanish production the feedback that they received from native speaker interlocutors. They used an L1 listening span task adapted from Daneman and Carpenter's (1980) sentence span test to obtain their WMC measure. Their WMC task required participants to listen to 36 sentences (3 sets of 3, 3 sets of 4, and 3 sets of 5) and while making

judgments on the sentences' grammaticality and plausibility, hold in memory the sentence-final words. Trained interlocutors then engaged in four tasks with the participants, which were recorded, and the participants' ability to modify their output based on the feedback they received was coded and rated. The results indicated that WMC did significantly predict learners' ability to notice recasts and corrections in the feedback that they received, leading the researchers to hypothesize that WMC does affect how language learners are able to attend to the details of language that they are receiving in real-time communication.

Kormos and Safar (2008) looked at PSTM and WM in relation to longitudinal development of L2 proficiency. Their study involved 121 high school students in Hungary who were learning English as a foreign language. They used a non-word span task and backward digit span task to estimate WMC and a non-word serial recognition test to estimate PSTM. They reported a lack of correlation for their PSTM and WMC measures (thus providing evidence that these two traits may be independent of each other) and that scores on both measures correlated with language development, though with different aspects of language development (hence their belief in two independent constructs). Their findings are interesting in that although they found a correlation between WM and PSTM and proficiency test scores, they did not see PSTM playing a role in language learning (as defined by improvement on the Cambridge First Certificate of English examination) for the lower proficiency group in their sample. This runs counter to some of the assumptions about WM, and PSTM in particular, regarding its being a mediating influence for beginners; but they were looking at all language skills, and perhaps their proficiency measure (Cambridge First Certificate of English) was at too

high a level to capture lower-end gains. Interestingly, in the correlations that they ran for all the language skills, listening correlated slightly higher with their WMC measures than did reading, speaking, or writing.

Hummel (2009) examined the role of phonological memory (PM) and aptitude for high proficiency speakers of English as a second language. Her participants were 77 native-speaking French university students who were studying to be teachers of English and had already completed seven years of English instruction. PM was measured by a non-word recognition task (non-words based on Arabic to reduce likelihood of advantages from phonotactic similarities with the L1 or L2). Multiple regression analysis revealed that both PM and aptitude were significant predictors of proficiency. Hummel then divided her sample to compare the higher proficiency half to the lower proficiency half, and contrary to Kormos and Safar (2008), found that PM and aptitude remained a significant factor for the lower group but not for the higher group (although the sample was initially described as quite homogenous, the standard of deviation for their L2 proficiency scores was quite high), again providing some evidence for WMC being more important for lower-level learners. French and O'Brien (2008) investigated the role of phonological memory and the acquisition of L2 grammar by French-speaking children who were learning English. They found that PSTM did contribute to the learning of new grammar structures, which corroborated L1 research findings about the role of PSTM in the acquisition of lexical items by children. They also found that participants' PSTM scores did not change over the course of the study, providing some evidence that PSTM ability may be a fixed trait.

Turning to listening, Miyake and Friedman (1998) conducted a study that looked into the possible correlations between WMC and the ability of Japanese learners of English to notice L2 linguistic cues and comprehend complex L2 syntax in aural stimuli. Regarding cues, Japanese speakers tend to rely on animacy or grammatical particles indicating topic or subject to provide them with information on the agent in a sentence, whereas English speakers rely overwhelmingly on word order. They designed their experiment to test the cue preferences and syntactic comprehension of 59 Japanese learners of English with sentences in English that were presented aurally. They also gave the participants listening span tests in Japanese and English to measure their WMC. In Japanese, participants had to recall the first word of several sets of sentences; in English the task was to recall the final word. So unlike the later Mackey et al. (2010) study, their listening span task involved the capacity aspect of WM, but not processing. Their path analysis indicated that L1 WMC determined L2 WMC and that L2 WMC determined participants' cue preference distance, and that both L2 WMC and cue preference distance determined syntactic comprehension. Also noteworthy is that although not explicitly stated in their hypotheses, the stimuli in this study were all presented aurally, so the question of listening comprehension ability is implicit in their questions of processing.

Another study that specifically looked at WMC and L2 listening was McDonald (2006), who reported on two experiments in which both non-native speakers (with a variety of L1s) and native speakers of English gave grammaticality judgments on spoken sentences. In one of the experiments, the native speakers were tested while having their WMC stressed by an additional demand (noise, time constraint, or digit load). The results indicated that the L2 learners performed more poorly than the unstressed L1s, that L2

WMC (measured using the L2) correlated with the L2 learners' ability to judge the grammaticality of spoken utterances, and that when L1 speakers had their working memory stressed, they performed similarly to the L2 learners.

Brunfaut and Révész (2011) looked at the impact of working memory and listening anxiety on listening task difficulty. Their study involved 93 students with a variety of L1 backgrounds at proficiency levels spanning primarily from A2 to B2 on the Common European Framework of Reference scale (Council of Europe, 2001). They measured working memory with a visual forward digit span test and a visual backward digit span test. The listening task had the participants listen to a short passage and complete 30 multiple-choice items requiring them to select the missing word (based on what they had heard in the passage). The researchers reported that neither participants' digit span nor backward digit span predicted listening scores. This finding runs counter to other research on WM and language processing, although as mentioned above, not much of that previous research has looked specifically at listening. On the other hand, the mixture of L1 backgrounds and proficiency levels in the Brunfaut and Révész study may be a factor. However, this possible limitation may need to be investigated further; if WMC is assumed to be a stable trait and WMC measures are language independent, it would seem to be permissible to involve participants from different L1s and proficiency levels in the same sample and still see an effect.

The research into WMC and L2 proficiency or L2 development has shown that greater WMC increases the likelihood of successful or more rapid acquisition of a variety of elements of the L2, but exploring the potential role of WMC in the process of understanding implicature has rarely been investigated. Taguchi (2008) is an exception.

Taguchi's earlier 2005 study included reflections on the possible influence of short-term memory constraints on her participant's propensity for choosing distractors that played off the last-heard utterance. Additionally, her 2005 results showed that proficiency did not influence speed of responding (reaction time measures for choosing a key) to her implicature items, and she wondered if individual differences might play a role. Her 2008 paper reported on an attempt to investigate some of the factors underlying the ability to comprehend conversational implicature in a second language, and included measures of working memory capacity, lexical access, and phonemic discrimination (in addition to an overall listening comprehension score and a pragmatic listening comprehension score). Her correlational analyses, however, revealed weak and non-significant correlations between her measure of working memory (a Japanese version of the reading span test) and both the overall listening score (an institutional TOEFL listening section) and the specifically pragmatic measure. However, the lack of a result in her study does not appear to provide sufficient evidence to abandon the notion that working memory contributes to successfully comprehending conversational implicature. Taguchi posits that perhaps the lack of a relationship in her study is because the working memory measure was taken in the L1 and not the L2. However, it is the author's view that a language-free working memory measure, which has been shown to correlate well with L2 listening ability in previous studies is perhaps a better way to get a clearer picture of whether or not greater working memory capacity is predictive of increased ability to comprehend conversational implicature. In Taguchi (2008) and earlier work on conversational implicature, there is talk about different degrees of implicature (high and low) and the differing numbers of contextual cues that contribute to the strength of the implicature. On the basis of this

consensus view (Taguchi, 2008, p. 520–521), there is no need to abandon the notion that working memory capacity plays a role without further examination.

Furthermore, beyond the narrow focus on conversational implicature, the continued effects of working memory that are seen for higher proficiency speakers of an L2 is an area worth investigating: In the words of Andringa, Olsthoorn, van Beuningen, Schoonen, & Hulstijn (2012, p. 53), "Exactly how working memory constrains comprehension is subject of debate." The notion that greater WMC provides a benefit for beginning learners is in accord with findings from L1 acquisition that see the strongest effects for WMC at the earlier stages of acquisition. One hypothesis about continuing effects of WM for higher-level learners is that it aids them in the ability to comprehend implicature. The research on WM and L2 proficiency is rich, although when looking specifically at WM and L2 listening, it grows scarcer, and when considering L2 listening comprehension of conversational implicatures, it grows even more scarce. Exploring the potential effect of working memory capacity on the process of comprehending conversational implicature in a second language is important and potentially beneficial, but this inevitably raises the question of the role of reasoning ability.

By being able to partial out how much variance is attributable to working memory, we will be able to gain insight that this cognitive factor is playing a role in the ability to carry out the parallel processes of extracting implicature from utterances that have been proposed by Wilson and Sperber. However, another piece of the cognitive puzzle is potentially found in reasoning ability. A number of researchers have looked into the possible link between inductive reasoning ability and L1 listening comprehension, and pursuing a similar line of investigation in L2 listening seems logical. In addition,

studies limited to one individual difference (if only working memory were included in the design) when investigating listening ability make the error of ignoring the multifaceted nature of both the listening process and of individual differences in general (Bodie et al., 2008, p. 111).

Therefore, it would seem that a measure of reasoning ability could potentially be informative. The rationale is that, if Wilson and Sperber and other researchers are accurate in their assertion that implicature is an extensive and common aspect of comprehension, it would stand to reason that reasoning ability is *not* likely to play a big role in first language listening. But for the L2 listener, it seems that the role of reasoning in comprehending implicature is an empirical question—and likely related to degree of strength or weakness of the implicature. Consider example (2.9):

(2.9) A: Have you seen John?
B: His car is parked outside Mary's house.

Assuming that A and B know each other well, and know John well, and that John is friends with Mary and that John rarely goes anywhere except in his car, A will extract the implicature that "I have not seen John but I saw his car outside Mary's house and therefore John is probably at Mary's house" on the basis of very little reasoning. The shared context precludes the need for reasoning. But let us consider example (2.10), taken from the 2013 movie *Silver Lining Playbook* in which a waitress says to a man in a diner:

(2.10) Waitress: Not so fast, raisin bran.

The implicature here might not be so readily available. However, based on the context that the man 1) ate a bowl of raisin bran in the diner, 2) did not pay his check, and 3) is on his way out the diner's door, the man, and any onlookers, or viewers of the movie

(after first potentially being confused by the address of "raisin bran"), will, after reflecting on the meal just eaten, draw the intended implicature from the waitress's utterance. So, degree of implicature is likely a factor. And by degree of implicature, we are talking about the number of cues provided to fill in the necessary information for the general social context, the specific social context and to identify the directly relevant factors. The greater the number of disparate cues (rather than those that are redundant or reinforcing), the more the impact of working memory capacity, rather than reasoning ability, will play a role. Therefore, measures of working memory are likely to be more informative than a measure of reasoning ability. Furthermore, the construct and the measures for reasoning are far fuzzier than they are for working memory capacity, and there are cognitive psychologists who discuss the similarities and overlap between the two constructs and assert that they are the same (Ackerman, Beirer, & Boyle, 2005; Engle, 2002; Kyllonen & Christal, 1990).

Chapter 3: The Current Study

While there are clearly a host of factors involved in any aspect of language use, including, as the literature review above shows, comprehension of conversational implicature, when discussing assessing that language use, test developers need to try to consider as many of those multiple factors as they can in terms of validity. Validity addresses the core question of trying to confirm that we are measuring what we have set out to measure. There is some debate about whether this is done by addressing the question of a valid test (Borsboom, Mellenbergh, & van Heerden, 2004; Borsboom, Cramer, Kievit, Scholten, & Franic, 2009; Lissitz & Samuelsen, 2007) or whether it is a matter of a valid score interpretation (Cronbach & Meehl, 1955; Chapelle, 2011; Kane, 1992; 2009). The focus in this study is the validity of a task and construct which potentially could be suited to either position, but for the sake of moving on to the question of the validity of the task and construct at hand, the view will be taken that the validity of listening scores (which is the prevailing view in the field) is the issue of concern.

The validity of a task, and therefore the validity of scores derived from performance on the task, can be bolstered by understanding the task. This is not groundbreaking, but it is surprisingly rare to see validity arguments for language tests made in conjunction with the actual content of the tests—this is likely the case because the test content changes from administration to administration, and there is an assumption that adherence to item specifications and pilot testing will smooth out the possible idiosyncrasies introduced by task content, but this is an assumption that ought to be investigated. This is therefore one goal of the study: to increase understanding of what is being tested in a particular test task, and therefore demonstrate the validity of its use. By

comparing the ability to respond to implicature items in a multiple-choice format and a constructed-response format, while also taking into consideration measures of working memory capacity and overall language proficiency, we can attain a better understanding of whether we are measuring what we set out to measure with conversational implicature items in a multiple-choice format. Messick's work (1989, 1994, 1995) highlighted the need and importance of different kinds of validity evidence and that different types of evidence help build the case for validity: "The varieties of evidence are not alternatives, but rather supplements to one another" (1989, p. 6). Messick was explicit that there are not different types of validity evidence that one can claim; i.e., that the validity question is a unitary one—and that researchers must bundle the different varieties of evidence that they have into making an overall argument. It therefore stands to reason that all else being equal, more varieties of evidence will lead to a stronger argument.

Kane's (1992, 2001, 2011) work on validity, in particular his notion of a chain of inferences, built on the work of Messick and provided a framework for language testers to more easily apply evidence to their claims. Figure 2 below, adapted from Chapelle's (2011) explanation of Kane's argument-based approach to validity demonstrates how this study can be framed as an investigation of validity.

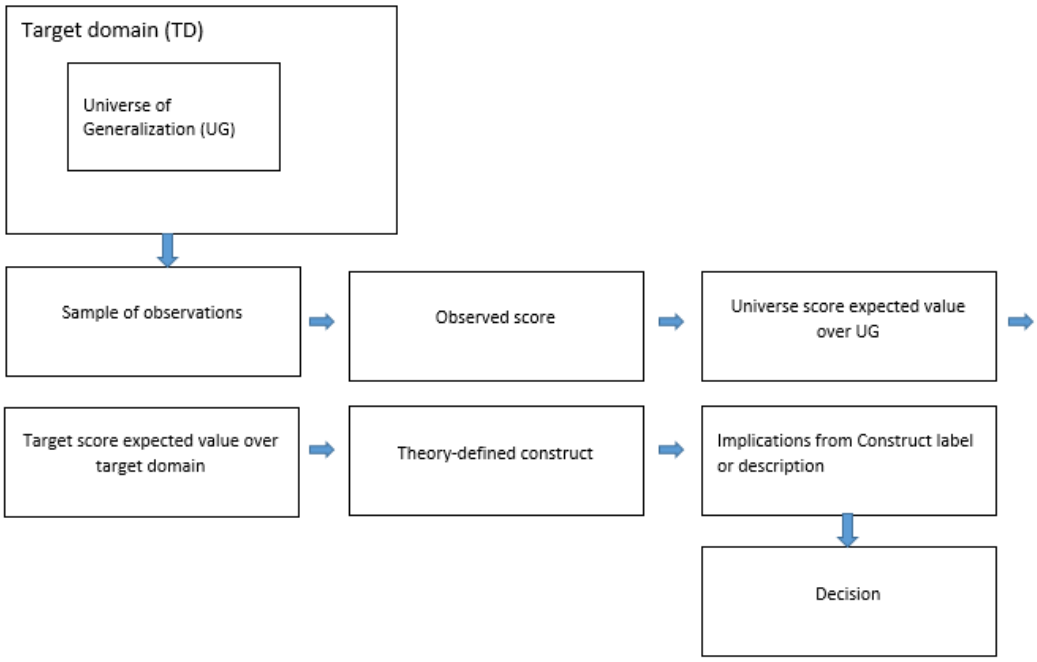


Figure 2. Interpretive Argument Chain

In the context of this study, we start with the (vast) target domain of "being able to interact competently in conversation." From this target domain the smaller subset *universe of generalization* is drawn, which in this case is "understanding conversational implicature." As discussed above, there is evidence that this universe of generalization is a subset of the target domain of interacting competently in conversation (i.e., demonstrating communicative competence in conversation). The next inference, from *sample of observations* to *observed score* is one of the links in the chain that this study is focused on. Current practice in standardized language testing involves the scoring of multiple-choice items for the assessment of comprehension of conversational implicature. But it is unclear how valid this practice is—that is, are the inferences that test takers are identifying when given a set of options the same inferences they would make without being provided a set of options. Comparison of constructed response difficulty estimates to multiple-choice difficulty estimates for the same items provides some evidence for or

against this assumption, along with an evaluation of the reliability estimates for the two formats in aggregate.

Additionally, moving three links up the chain in the figure to the inference from *target score expected value over target domain* to *theory-defined construct*, it is believed that examining performance on conversational implicature items in conjunction with learners' estimated overall language proficiency and estimated capacity for working memory will provide support for the theory-defined construct of implicature as a combination of higher-order thinking and language processing. Finally, the data to be gathered will also potentially provide meaningful information to help inform the final inference in Kane's chain, which links *implications from construct label or description* to *decision*. Analysis of participants' scores on a reliable measure of conversational implicature comprehension with their performance on a CEFR-linked measure of language proficiency will provide support for (or detract from) claims that implicature is a skill that learners should only be expected to have control over at higher (i.e., C1 on the CEFR) levels of proficiency—independent of their working memory capacity. In regard to the CEFR, while it has its share of detractors, the idea of a framework of reference that increases interpretability of test scores across languages and contexts is appealing, which is likely why the CEFR has spread far beyond Europe since it was introduced: it is extremely useful (to learners, employers, etc.) to have a way to indicate that a Spanish L1 learner with a score of X on a German proficiency test and a score of Y on an English proficiency test has comparable skills in those two languages, or is much stronger in one than the other. Some researchers (e.g., Fulcher, 2004) have questioned whether this is possible (or desirable); others have been less skeptical (Alderson, 2002), while still

expressing the caveat that such linkage is only possible if the tests are developed from the design stage with the CEFR in mind. The CEFR manual for linking itself recommends taking a conservative approach to such endeavors (Council of Europe, 2009, pp. 1–2). But it should be pointed out that if a test is created with the CEFR as the basis for test specifications, rather than successful language use in a particular language-use context as the starting point, it is difficult to counter the possibility of circularity when that test is linked to the CEFR. On the other hand, if Test A is designed to assess a particular language-use ability in a particular context, it is possible after the fact, through a formal linking study, to examine the content of the test in relation to an external framework such as the CEFR to establish how test-taker performance on Test A relates to that framework.

However, linkage to the CEFR is possible whether the CEFR was involved in the test design stage or not, on the basis that the progression of ability laid out in the descriptors in the CEFR's 53 illustrative scales have an empirical grounding in reality. This empirical grounding is described in North's (2000) initial scaling work with the CEFR's descriptors, which has been replicated by other researchers (e.g., Alderson, 2005: Kaftandjieva & Takala, 2002; Papageorgiou, 2009). However, from the original document to more recent writing on the CEFR by one of its primary authors (North, 2014), some consistent themes of discussion surrounding the CEFR have been that 1) the CEFR is not a test blueprint, but rather a framework to be used as a point of orientation, and that 2) the CEFR is a work in progress, in need of further specification. It is in regard to the second point that the inclusion of learner CEFR levels as a variable in this study is of value; because when one examines the listening scales of the CEFR closely, the potential under-specification of understanding implicit meaning becomes apparent.

However, it also bears mentioning that the second point is only an issue if the CEFR is not being used as a framework and point of reference as originally intended. As a framework, it should not need further specification. The reality, however, is that its use has gone beyond being a framework, and therefore the question of under-specification is often central.

Turning to some of the document's scales themselves (n.b., some CEFR scales were revised slightly with the release of the 2018 companion volume (Council of Europe, 2018), so some of the scales discussed here are from the original 2001 publication and others from the 2018 revision), when one reads the Overall Listening Comprehension scale (Council of Europe, 2018, p. 55), we see that for the C1 level, it includes the descriptor, "[c]an recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts," and "[c]an follow extended speech even when it is not clearly structured and when relationships are only implied and not signaled explicitly." The B2 descriptors make no mention of grasping implied meanings (whether in extended speech or not), but it is strongly implied when reading that at the B1 level, learners are only at the point, given their limited vocabulary and structural knowledge, of being able to comprehend main ideas and some details from straightforward factual input. So there is a jump in expected type of comprehension ability from B1 to C1, without addressing clearly what might be expected of a B2 level listener with non-factual, non-straightforward input—it is only the complexity and structure of the input that is specified at B2. Whether "complexity" extends beyond syntactic structures or lower frequency vocabulary is unclear.

On the other hand, the scale pertaining to Understanding Conversation between Other Speakers (Council of Europe, 2018, p. 56) moves from lower B2 learners requiring effort towards (but being capable of) understanding what is being said by other speakers conversing with each other, to higher B2 learners being able to comprehend that type of input even if the speakers are conversing without adjusting their speech rate or lexical choices (i.e., "animated") out of consideration for less proficient speakers. Based on the frequency with which language users employ implicature, this would strongly imply that at the B2 level, this ability is something that the language learner has developed and is able to use in communicative situations.

The scale describing ability to identify cues and infer (Council of Europe, 2001, p. 72), again makes explicit that at the C1 level, the ability to use context and linguistic cues to infer is expected, but in its descriptor of the B2 level, it leaves open the door to the ability of learners at that level to handle this type of input with the ambiguous reference to "variety of strategies" employed to make use of "contextual cues." Finally, the Sociolinguistic Appropriateness scale (Council of Europe, 2001, p. 122) again indicates an expectation of the C1-level learner being able to readily attain intended implicatures (e.g., "allusive and joking usage"). But in the descriptions of B2, test developers and other test stake-holders must again make larger inferences to reach the assumption that learners at this level are consistently comprehending implicatures correctly (i.e., keeping up with and contributing to group discussions even when speech is fast and colloquial, and not interrupting the flow of the communicative interaction with errors or miscomprehensions).

Tying functions, or even mastery of functions, to proficiency levels is not the goal of the present study. Specific language functions (complaining, apologizing, etc.) are not what is being assessed, rather it is when these functions are performed via conversational implicature. Further, as the multiple-choice task type under investigation here is one that is widely used in language proficiency tests, trying to determine if the implicature is the source of difficulty rather than the linguistic input itself is of value to language testers, who continue to rely on imperfect information that implicature items are "more difficult" than non-implicature items. Therefore, it is an assessment issue rather than a language development issue (as is often the case whenever language proficiency descriptor scales are in use).

But because lack of specification of the CEFR has been a widespread criticism since its inception—the framework creators themselves repeatedly refer to the CEFR as a document in progress and as a document to use as a point of orientation for teachers and testers rather than as a prescriptive tool—an empirical basis to link the understanding of conversational implicature to a CEFR level would be of benefit to the field: what Alderson et al. wrote in 2006—"an urgent need exists to illustrate the levels of the CEFR with calibrated test items" (p. 4)—still seems to stand.

To summarize the gaps in the literature that this study tries to address, from work in pragmatics we know that conversational implicature is a frequent aspect of communication; therefore, it is a necessary skill for language users to have. But while we understand and know that different types of inferencing contribute to difficulty and tax processing more than when not present, there is a lack of clarity on whether this type of listening is actually separate from other types of listening. A possible reason for this lack

of clarity is that there has not been a balanced approach to looking at this type of listening with an instrument akin to many standardized listening instruments. Understanding the separability of comprehension of conversational implicature better, with different sources of evidence stemming from different analyses, will provide insight on how great the need is to specifically target this subskill in listening comprehension tests. Towards that goal, answers to the research questions listed in section 3.1 below were sought.

3.1 Research questions

1. On the constructed response version of the items, will second-language learners be able to generate the expected inferences that experienced test developers assume they can?

2. To what extent do items targeting comprehension of conversational implicature and items targeting general understanding test the same construct?

3. Will between-subject performance on conversational implicature items vary significantly at the B1, B2, and C1 levels (as estimated by a CEFR-linked measure of proficiency, e.g., Michigan Language Assessment's MET)?

The fourth research question centered on the possible influence of short-term memory and working memory, and was subdivided into three questions:

4a. To what degree does the role of working memory (as measured by both complex and simple tasks) influence performance when ability to comprehend conversational implicature is measured by multiple-choice items?

4b. To what degree does the role of working memory (as measured by both complex and simple tasks) influence performance when ability to comprehend implicature is measured by constructed-response items?

4c. To what extent will working memory exert greater predictive influence on performance for multiple-choice items in comparison to constructed response items?

The prediction for the first research question was that with an assumption of test-taker's general world knowledge and through the use of context cues in the input, target ability-level test-takers would be able to generate expected keys in the constructed-response format that match the keys that are provided in a selected-response format like multiple-choice items. This prediction was borne out and will be discussed in section 5.3 below.

The prediction for the second research question, regarding evidence of a distinct listening for conversational implicature subskill, was that the confirmatory factor analysis model that includes listening for implicature and listening for general understanding as distinct factors would provide a better model fit than a single factor listening proficiency model. This would provide evidence that understanding implicature is a separate trait from general listening ability. This prediction was not borne out in the data; the single factor model provided a better fit, as will be discussed in section 5.4.

For the third research question, the prediction was that a meaningful difference in ability between CEFR B1 and B2 levels would be seen, but such a meaningful difference would not be evident between CEFR B2- and C1-level participants. Although it may be expected that different proficiency levels will have different means on the implicature items, information in this regard has been lacking in terms of the CEFR and is of interest to test developers who use the CEFR scales as point of orientation for estimating test levels. The results, however, showed that the difference between C1- and B2-level participants, as implied by the CEFR scales, was observed; the predicted difference

between B1- and B2-level participants was not. The analyses on this question are discussed in section 5.5.

For the research questions focusing on the memory measures, the predictions for research questions 4a and 4b were that the null hypothesis would be rejected. The prediction for research question 4c was that the difference would be non-significant (i.e., working memory plays a beneficial role in comprehending the implicature, but test method will not have an effect). As will be discussed in detail below in section 5.6, the short-term memory measure results could not be analyzed, but when examining the role of working memory, a small role was seen with the general, non-implicature items, but not with the conversational implicature items as predicted.

Chapter 4: Method

Answering the research questions in this study regarding the underlying construct of a listening test and the potential role of individual factors such as English language proficiency and working memory on performance with the listening test required an experimental methodology and the use of multiple statistical analyses. The analyses that were used are described in detail in Chapter 5, but to explain the study's sample size it is necessary to mention two of them briefly: the Rasch analyses and the confirmatory factor analyses. The Rasch analyses, which were used to score the listening test and evaluate the performance of the listening items, and the confirmatory factor analyses (CFA), which were used to examine the construct of the listening test, are both statistical methods for which at least 200 participants is suggested (Linacre, 1994 for Rasch; Brown, 2006 for CFA). It is for this reason that a sample of 250 participants was sought. Additionally, due to the experimental design and involvement of human participants, Institutional Review Board (IRB) approval was sought and attained from the University of Maryland IRB (Project 873347-1).

4.1 Participants

Participants for the study were recruited from two language learning organizations in two South American cities in June and July 2017 using IRB-approved emails and fliers. A total of 255 Spanish-speaking participants enrolled in the study: 200 in four days at the first location and 55 in two days at the second location. All who enrolled in the study read and signed Spanish-language translations of the consent form and all were paid $10 US for their participation. The compensation aspect of the study was not emphasized at either location where the study took place and therefore it is deemed unlikely that payment had any meaningful impact on their performance and thus the data. As became clear while

proctoring the sessions, the primary draw for some participants was the fact that the listening test could serve as "practice" for one section of the Michigan English Test (MET) listening test. The MET was in use as a proficiency test at both locations and some participants were eager to receive raw scores of their performance on the experimental test. The option to receive raw scores was explained to them in the study session introduction; participants simply had to enter an email address in the final "comment" field that was on the last page of the listening test; overall, 40 of the 255 participants (about 15%) asked for and received their listening test raw scores.

Of the 255 participants who enrolled, four did not consistently use the unique identifying code they were given and were excluded from the study (all four from the first location), leaving a final sample of 251. Table 2 provides a summary of participant demographics for each location and overall. This information was collected via a short questionnaire that preceded the listening test (see Appendix C for demographic questions).

Table 2. Participant Demographic Information

	Male / Female	**Average age (S.D., median)**	**Average years spent in English-speaking country (S.D., median)**
Location 1	89 / 105*	22.3 (5.38, 21)	0.38 (1.6, 0)
Location 2	17 / 38	27.5 (8.64, 26)	0.45 (1.12, 0)
Total	**106 / 143***	**23.6 (6.78, 21)**	**0.39 (1.47, 0)****

*No response = 2, ** No response = 22

Per the parameters of the IRB approval, the minimum age for participant was 18; no maximum age was set (the oldest participant was 56). All participants were recruited from within the two language learning institutions and the vast majority were students in

those institutions. (In location 1 all participants were students; in location 2 some participants were teachers or staff.) The sole criteria for enrollment beyond a minimum age of 18 was intermediate or high-intermediate proficiency in English/English listening. (Level of proficiency was estimated by staff at the institutions; there was not a proficiency screener.) Information on years spent living in an English-speaking country was collected as a confirmatory check on the learner status of participants. The values entered for time spent in English-dominant countries were uniformly small, with 84% of the participants (147 of 174 providing a response) from location 1 and 78% of the participants (43 of 55) from location 2 reporting 0 years (i.e., 6 months or fewer). Only two of the participants (both from location 1) reported having spent more than 10 years in an English-speaking country. One participant who was 18 at the time of the study indicated 10 years (with initial exposure to English at age 7), and the other who was 30 at the time of the study indicated 15 years (with an initial exposure to English at age 14). Although there is some potential that both could be defined as bilingual from childhood, both reported Spanish as their first language and both were residing in a country where Spanish is the predominant language at the time of the study. With that in mind, and because the study was not looking at the effect of English exposure or any questions around sensitive periods, their data was not removed.

As noted above, both institutions where the study was conducted administer and use the Michigan English Test (MET), which is an A2–C1 general listening and reading proficiency exam (Michigan Language Assessments, 2019). Because the exam is used widely at both institutions, efforts were made to recruit from learners who had taken the MET or who had been placed in classes on the basis of their MET results. On the day of

the study participants were asked to report their MET listening scores in the background survey to provide an independent measure of proficiency. As will be discussed in detail in section 5.5 below, 84 of the 251 participants were able to reliably provide their MET listening scores.

4.2 Instruments

Five instruments were administered to the study participants: a test of conversational implicature, two working memory measures, and two short-term memory measures. Information from a sixth instrument, the MET, was requested in order to have an independent measure of English language proficiency.

4.2.1 Measure of conversational implicature listening ability

A test of comprehension of conversational implicature (TCI) was created for this study. As mentioned above, the reason for creating the TCI rather than using existing standardized test materials is that one of the limitations of past explorations of listening subskills is their reliance on test items which were coded for subskill post-hoc or which did not contain equal numbers of items of the subskill(s) of interest. The TCI included 30 items designed to assess conversational implicature and 30 items designed to assess "general" comprehension of explicit information. All 60 items were created by the author, who at the time of the test's development had more than ten years of experience creating multiple-choice listening comprehension items for high-stakes assessments.

All items were created with the same specifications (word count, number of speakers, number of turns, etc.) except for target subskill of "implicature" or "general" listening (see Appendix A for item content and Appendix B for item metadata). The development of the 60 test items in the TCI followed industry-standard methods for creating listening comprehension multiple-choice items. After the items were drafted,

proofread, and checked for vocabulary use at similar levels of frequency, audio files were created. All 60 item dialogues were recorded by the same adult male North American English L1 speaker (the author) and adult female North American English L1 speaker. The directions and item stems were recorded by a second adult male North American English L1 speaker. Because the author was directly involved in the recordings, it was possible to ensure that the prosody of the speakers matched the intent of the items. This was accomplished by doing an unrecorded dry-run of each item first, confirming that both speakers were delivering their turns with the appropriate prosody and then making the recording. Having test developers act as "directors" (or providing clear prosodic information in the scripts for voice talent) is an important aspect of audio recording for all listening comprehension tasks, but is especially important when the testing focus is conversational implicature. Audio files were created for each step of the review process as they are necessary for reviewing listening items in a theoretically sound manner: a reviewer's first exposure to listening stimuli should be through the auditory mode and the best qualitative check of whether an intended meaning that is being tested is accessible to listeners is to verify it through reviewers generating the keys themselves based on the audio stimulus and stem alone.

Four experienced test developers in addition to the author were involved in the three rounds of review that the items underwent. The first drafts of the items were given an initial content review by an experienced test developer, who was provided the audio files and a Word document containing the items and item metadata. When the review was completed, feedback was sent back to the author in the form of comments in the Word document. Items were then revised as necessary and the revised items (with revised audio

if necessary) were sent to a second equally experienced test developer in the same manner as the first (although later reviewers did not have access to the comments and feedback from earlier reviewers). Feedback from this second content review was incorporated into the items and a final content review was made by a third experienced test developer. This method of item development—multiple stages of review in addition to the original writer—is the norm (AERA, APA, & NCME, 1999) for high-stakes standardized testing. The constructed response versions of the items were created by simply replacing the multiple-choice options from the finalized MC versions with an open text field in which the participants produced their response to the question.

Although all implicature items were coded for implicature type (i.e., relevance, indirect criticism, irony), the degree of implicature required was given more consideration. The reason for this is that past work (Roever, 2013; Taguchi, 2005, 2009) has already fairly clearly established that conventional implicature (e.g., indirect refusals, indirect requests) are easier for L2 learners than conversational implicature. Additionally, within conversational implicature, because the role of context is so important, this is an area that test developers wrestle with: testing implicature as genuinely as possible (i.e., with less context explicitly provided, as is the norm for two speakers who work with or know each other) while also allowing for stimuli to be accessible. Therefore, the implicature items were developed to try to cover three levels of strength of implicature: weak, medium, and strong. Following Kasper (1984) and Taguchi (2008), evaluation of the degree of implicature was based on whether the general context is provided (or whether the listener must infer it), whether the specific context is provided (or whether the listener must infer it), and the number of relevant disparate indicators (or cues)

provided that the listener must process in order to reach the tested implicature. Weak implicature items are ones where listeners are required to do a high degree of inferring; that is, indication of context and number of cues are minimal. Strong implicature items are ones where listeners' implicature requirement is low; that is, context and/or multiple cues point to one logical or reasonable meaning. (It should be noted that the items were developed with only an experienced item writer's intuition about the difficulty of different types of implicature and the systematic coding was applied afterwards. It is for this reason that the question of degree of implicature is not the focus of any research questions (although the results of item performance by degree of implicature are reported in Section 5.3), but a consideration of the concept is essential for test developers who are creating implicature items and task; the question of the value of pursuing this area further empirically will be addressed in the Chapter 6 discussion.)

Example (4.1) shows a strong implicature item that was included (LI_01) in this study:

(4.1) W: How was your trip to New York?
M: First my car got hit when I left it parked overnight on the street, and then my wallet was stolen.

Question: How does the man feel about his trip to New York?

The rationale for a judgment of strongly provided implicature is that even though the general context is not provided (this can be conceived of as the domain—academic, occupational, personal—within which the exchange is occurring), it is not relevant. The specific context, or topic (the man's trip), is provided, and while there are two cues for the listener to catch and interpret ("car got hit," "wallet was stolen"), they are somewhat redundant in that they are both quite negative. The result is that the vast majority of

proficient and reasonable language users would agree that the implicature is that the trip to New York was not enjoyable for the man; this is what makes the implicature strong—alternate implicatures are not readily available to the listener and the degree of implicature required by the listener is low.

Example (4.2) shows an item (LI_02) that was coded for medium strength of implicature.

(4.2) M: Wow, it's really beautiful outside. Not a great day to be cooped up inside.
W: Yeah, it is. Let's go grab a cup of coffee, take a walk around the block or something.
M: Sounds good, but I've got a ton of work to do.
W: Suit yourself. I'm going to go stretch my legs.

Question: What will the man do?

Here the general context (workplace) is not provided, but it is somewhat relevant. Nor is a specific context (topic)—taking a break—provided. Although there is only one cue for the listener to interpret ("ton of work to do"), the lack of general and specific context should make this a more difficult implicature to reach than in (4.1).

Example (4.3) shows an item (LI_15) with a implicature coding strength of weak:

(4.3) M: I can't believe the way the consultants submitted the data. Information repeated in different places with different labels, inconsistent coding…
W: So, send it back. We're paying them!
M: I know. But it'll take me at least an hour to spell out everything that's wrong with it and <u>then</u> it'll take them another couple weeks to get it fixed…
W: Oh, no. I think I know where this is going.

Question: Why is the woman worried?

In this item, the general context (occupational, office setting) is not provided. A specific context (topic) is provided: the poorly prepared data, but there are numerous disparate

cues to process and interpret in sequence ("repeated information," "inconsistent coding," "we're paying them," "it'll take me at least an hour" "another couple of weeks"). But what primarily leads this item to be classified as a weak implicature is that there are potentially more than one defensible implicatures available to the listener. That is, the woman might be worried because she thinks the clean-up task will be assigned to her or she might be worried because the consultants' poor job has caused serious harm to their timeline. In multiple-choice test development, efforts are made to avoid using items with multiple available implicatures, unless items of greater difficulty are needed, in which case efforts are made to test the "majority" or "most tip of tongue" response and avoid including alternate implicatures in distractors.

As noted above, the degree of implicature evaluations were made solely by the author, and while two other reviewers evaluated the degree of implicature tags as part of their review, they were not asked to make independent judgments. However, in order to verify that the coding of degrees of implicature has some basis in reality, and that the items did not contain basic structural problems (double keys, confusingly worded stems, etc.) that the reviewers overlooked, two small-scale pilot tests of the first round of items developed (34 total, 24 implicature and 8 general) were conducted. One pilot was with native speakers of English (n=17) and one with non-native speakers of English (n=38) clustered around the target level of proficiency (high intermediate) for the study. All items were piloted through LimeSurvey, an open-source online survey tool. After being given directions and seeing a sample item, the test was delivered one item at a time. The audio began to play automatically when participants moved to a new page (with items preceded each time by the direction "Listen to the conversation and then answer the

question" so as to give participants 2–3 seconds to acclimate to the page before hearing the exchange). The audio could not be paused or re-played. Figure 3 shows how the multiple-choice items were presented on screen and Figure 4 shows the constructed-response format. In both formats the stems (and options for MC format) were visible while the dialogue was playing.

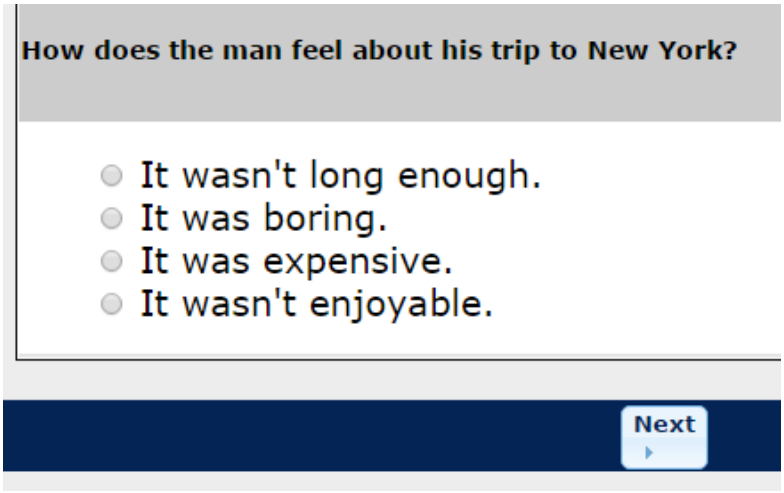


Figure 3. Multiple-choice Format Item in Test of Conversational Implicature

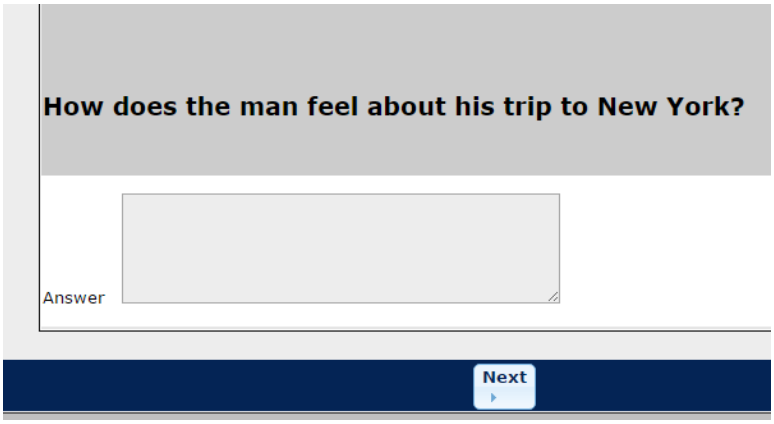


Figure 4. Constructed-response Format Item in Test of Conversational Implicature

There were two reasons for the interest in contrasting multiple-choice item performance with constructed-response item performance. First, as mentioned above, much research has been conducted on the validity of multiple-choice items, to counter

claims that test-takers are 1) only able to respond correctly to questions because answer options are provided and 2) that items are answerable without reference to stimuli. While research on these questions has confirmed that multiple-choice items are a valid method for testing comprehension (Freedle & Kostin, 1996; 1999; Kostin, 2004), the literature review did not indicate this question being looked at in a focused manner with implicature items, where the question of multiple correct answers is an intuitively larger issue than for more straightforward, factual types of comprehension. The second reason to be interested in looking at learner constructed-response answers is because with advances in machine scoring (Carr & Kunnan, 2016; Crossley, 2016), the likelihood of short-answer constructed response items appearing on large-scale standardized tests in the near future to test a variety of comprehension abilities is becoming more realistic. An exploration of performance on implicature items in the traditional MC format in comparison to a CR format could provide evidence to support this endeavor (as would a post-hoc analyses of the variety and types of responses produced in the CR format, although that is not the focus of the current study).

The pilot of the TCI with native speakers, who had varying levels of education and spoke at least two distinct varieties of English (Irish English and American English), was conducted solely with the constructed-response format of the items. This was done in order to verify that the expected responses (and thus the "answers") generated by the test developers would match that of native speakers who are not test developers. The same 32 items (all created and reviewed during the first round of item development) were administered to all participants in two orders to counteract any potential effects of order. The participants' (8 female, 9 male; average age 36.8 years old) responses to the items

were reviewed by the author to verify that they matched the key that the test developers assumed was the tip-of-tongue—or was at least the most plausible—implicature derived from the exchange. The raw score and percentile correct information of the English native speaker pilot showed that a great majority of the implicatures being assessed in these items were readily available to the vast majority of the participants. However, these results also show that items that were designed with their degree of implicature judged to be "high" were more difficult than those pre-judged to be low or medium, and that at least two of the items (LIC_03 and LIC_15) merited closer inspection (both were revised and their CR version difficulty estimates showed moderate difficulty). These results also gave some indication that an attempt to finely grade the difficulty of implicature items beyond a dichotomous level might not be quantifiable (i.e., medium and low degree might not be borne out in the data as the native speakers performed near ceiling on all low degree and medium degree items). As for native speaker performance on the general, non-implicature items, they performed at or near ceiling on seven of the eight items. While it would seem intuitive that native speakers would perform at ceiling on a language proficiency listening test, the check was necessary as some research, for example, Campbell, S. G., Hughes, M. M., Smith B. K., Meyers, J. H., O'Connell, S. (2012), has shown that if listening is designed to be challenging, as was the case in the Campbell et al. study, which aimed for a level akin to C2 on the CEFR, it will show variation in native speakers and the goal of the listening instrument in this study was to not be at that level of difficulty.

A small pilot with non-native speakers of English was also conducted with the items from round one of development (32 total, 24 implicature and 8 general). This pilot

included both MC and CR versions of the items. In order to ensure that each item received exposure in two different positions of the test in both its multiple-choice and constructed-response format, four forms of the test of conversational implicature were created for the non-native speaker pilot. The sample size was too small for conducting Rasch analyses, but looking at the items' facility and discrimination values, only four (LIM_03, LIM_04, LIM_05, and LIM_06) of the 32 items were flagged for problematic discrimination in their multiple-choice format and all four were revised prior to the full study (and examination of Rasch statistics for those items after the full study showed all with adequate fit). The reliability values estimated for the pilot forms was mixed (sub-par .501 for Forms 1 and 3 combined but an adequate .806 for Forms 2 and 4 combined), but because Forms 1 and 3 had an even smaller sample size (n=17) than forms 2 and 4 (N=21), and because all four poor performing items were on Forms 1 & 3, it was believed that revisions of the items would result in improved reliability—and that was the result (described in section 5.1 below). Essentially the small-scale pilot of the round 1 items showed that the expertise and multiple reviews of the item developers did result in a large majority of items that "worked" for English L2 learners at the target level.

After pilot 1, it was decided to increase the total number of items to 60 (from 34) and to ensure that there was an equal number of items for both the implicature subskill and general listening subskill. This meant that in round 2 of item development an additional 6 implicature items were created, so there were 30 total, and an additional 22 general listening items were created, so there were 30 general items as well. This allowed for the study to more clearly address its goal of looking at the implicature subskill in contrast to non-implicature listening, and also avoided the limitation of some past

listening subskill analyses that had unbalanced numbers of item by subskill type. The additional 28 items (6 implicature and 22 general) were all developed with exactly the same procedures as the original 32 items (all created by the author and reviewed with audio by three separate experienced test developers), but because of logistical and schedule constraints, these items were not piloted. However; performance of Round 2 development items did not differ in any meaningful way than performance of Round 1 development items.

After all items were finalized and re-recorded (if necessary), the 60 items were assigned across seven 40-item forms within LimeSurvey to allow for all 60 items to be administered at least twice in both formats and to ensure connectivity of the data for the Rasch analyses (Appendix D provides details of item placement across the seven forms).

The participants moved through the forty-item test at their own pace. Because the goal of the study was to examine comprehension of input without the pressure of a clock, time limits were not included. Participants completed a block of twenty MC items first and then moved to a twenty-item block of CR items (both preceded by format-specific directions and example), or depending on which form they were assigned, completed their twenty-item CR block first and then moved to the twenty-item MC block.

4.2.2 Measures of working memory capacity

The WMC of participants was measured using the Blockspan and Shapebuilder tasks, both of which are visual-spatial tasks. These two tasks were developed by Michael Dougherty, professor of psychology at the University of Maryland. The Blockspan task requires participants to observe a series of flashing lights on a grid (Figure 5) and then recall them in the correct order and location. Participants respond to 16 "rounds" of stimuli in the Blockspan task. The items increase in difficulty as a result of the number of

flashing lights in a series increasing as you progress through the task and as a result of a black "masking screen" appearing between some of the lights flashing in the later series of items. Participants receive scores based on how well they are able to identify the correct order and locations of the flashing lights immediately after seeing them. They provide their answers by clicking on the blue squares where they saw the yellow squares appear after being prompted with the green "Go" direction.

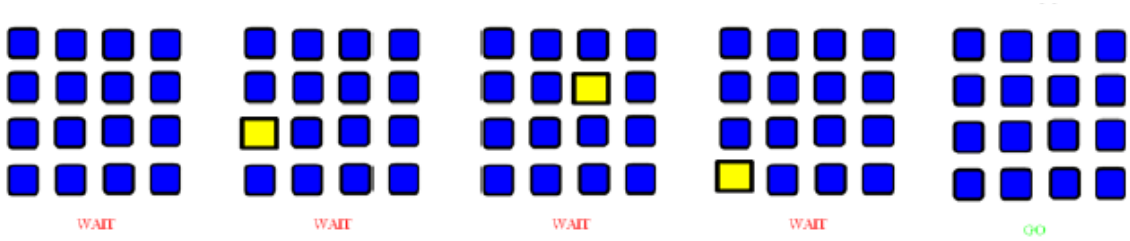
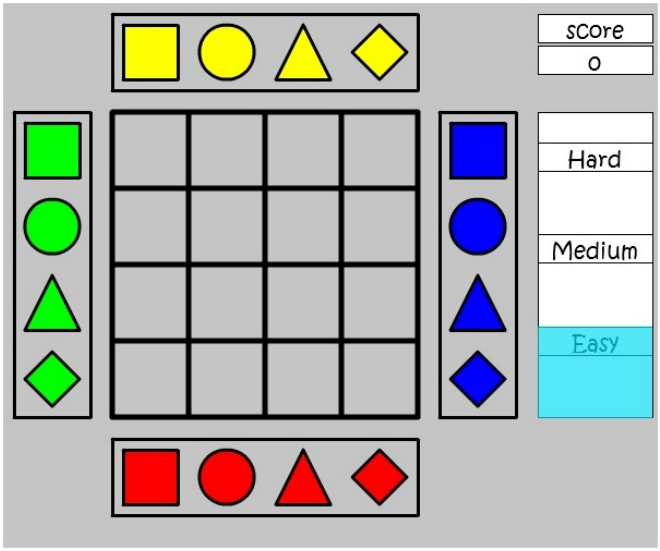


Figure 5. Linear Representation of Blockspan Three-Block Item

The Shapebuilder task requires participants to observe a pattern of multi-colored objects on a grid (Figure 6) and then recall the correct shapes, colors, and locations of the objects. The Shapebuilder task requires participants to respond to 26 rounds of stimuli (which increase in difficulty by increasing the number of objects to remember) and scores them based on how many correct responses they give. They provide their answers by dragging and dropping the shapes from the rectangles on the edges of the grid to the cells in the grid where they observed a particular color and shape combination.



-

Figure 6. Shapebuilder Screen Display

As mentioned in Chapter 2, some of the issues with past research into WMC and L1 and L2 abilities has centered on the potential confound between the WMC measures and existing proficiency. That is, when L2 learners are giving cognitive measures in a second language, however simplified that language is, the learners' level of proficiency in the language will likely have a moderating effect on the results of the cognitive measure. Using a measure of executive control that does not involve language is one way to avoid this issue. It is a widely held view in the literature on working memory that the executive control component of working memory "is domain general, so performance on spatial tasks should be relevant to verbal complex tasks" (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle (2005, p. 771). Measures such as Blockspan and Shapebuilder, which tap into both storage and processing abilities, are seen to be better predictors of language processing ability than those that only measure capacity (Daneman & Merikle, 1996). The Blockspan and Shapebuilder tasks have been used in numerous studies and have consistently demonstrated high reliability and a correlation with second language

proficiency (e.g., Clark, Wayland, Osthus, Brown, & Castle, 2013; Nielson, 2014). Because the web-delivery platform for Shapebuilder and Blockspan were only provided in English, a Spanish translation was created prior to the study and was provided to all participants as a handout immediately prior to their starting the tasks.

4.2.3 Short-term memory measures

Two measures of short-term memory, or simple working memory, were also included: the forward auditory digit span task and backward auditory digit span task. Research on L2 development and working memory indicates that it is likely that complex working memory tasks are stronger predictors of L2 outcomes (Ackerman et al., 2005; Juffs & Harrington, 2011; Linck et al., 2013) than simple working memory measures. But despite the tendency of simple working memory measures to not correlate as strongly with complex cognitive tasks (such as listening comprehension) as working memory tasks do, whether one is conceiving information processing in terms of the Baddeley or Cowan models, or earlier working memory models (Broadbent, 1958; Atkinson and Schiffrin 1968, as cited in Cowan, 2005), the notion of a short-term memory span is part of the construct. Cowan (2005) summarizes the core commonality of all these models as being an explanation of the transfer of information (aural input in the case of this study) from 1) a passive but unlimited capacity short-term memory to 2) the limited capacity "processing" (complex WM) to 3) long-term memory (although whether and how input from STM interacts with LTM before entering WM is an open question). The overlap between STM and measures that also include executive function processing is an overlap of the memory process (Engle, 2002, p. 21). Therefore, by including both types of measure and examining the differential impact each has on performance on a complex listening task, and by examining shared variance, it should be possible to see how much

additional variance is explained by the "processing," or executive function, aspect of the complex tasks. Furthermore, some studies looking at L2 proficiency that have included STM measures in their models have found predictive influence. Kormos and Saffar (2008), for example, found that 30.25% of variance in English proficiency scores was explained by their backward digit span task; Andringa et al. (2012) found that two of five simple memory tasks correlated positively with listening comprehension performance, and Linck et al. (2013) noted that phonological short-term memory was found to be one of the strongest predictors of attaining high level proficiency in a second language. Essentially, as Juffs and Harrington wrote in their review of L2 and working memory, "the relationship between the two [short-term memory and working memory] remains an open empirical question" (p. 141), and following Linck et al.'s (2014) recommendation to include both types of measures until a clearer picture is formed, short-term memory measures were included to provide the possibility of comparing performance on them with the WMC tasks.

The forward and backward auditory digit span tasks used in this study were created by the author in order to be able to deliver them through a web-based platform (i.e., LimeSurvey). Existing FDS and BDS measures were available but required the ability to download an application to computers that was not possible in the context of data collection for this study. As is the norm with digit span measures, the items were presented in sets of increasing length (Juffs & Harrington, 2011). Specifically, the measures were created to deliver four 3-, 4-, 5-, 6-, and 7-element strings of numbers for a total of twenty items in each measure and were delivered in order according to length. The items consisted of a recording of a male North American English native speaker

uttering the numbers (in English) with a 1.5 second pause in between numbers. Each item was played only once. Directions were provided prior to the task on the screen in both English and Spanish and participants were given a warning on the screen when item length was moving from three to four to five, etc. Participants responded to each item after the recording played by typing the numbers they heard into a text field. Three versions of both the BDS and FDS tasks were created in order to minimize possible effects of item order.

4.2.4 Independent measure of English language proficiency

In order to address one of the goals of this study—provide a correlate of proficiency regarding comprehending implicature for the CEFR—a measure of participants' English comprehension listening ability that is separate from the measure being constructed for this test and which is linked to the CEFR was required. The MET, a general English proficiency listening and reading test which has been formally linked to the CEFR (Cambridge Michigan Language Assessments, 2012), was selected for this purpose. While only 84 of the 255 participants provided their MET scores (with an additional 29 participants reporting that they could not remember or had not yet received their scores), the analyses necessary for answering this question still could be conducted on that subset of the sample and are described in section 5.5.

4.3 Procedure

Staff at both locations where the study was conducted assisted with setting up data collection days and times. In location 1, four days with four sessions (8–10am, 11am–1pm, 3–5pm, and 6–8pm) on each day were set up, and at location 2, two days with four sessions (8:30–10:30am, 11:30am–1:30pm, 3:30–5:30pm, and 6:30–8:30pm) on each day were set up. Participants were recruited for the sessions in the 1–3 weeks prior to the

sessions via emails or fliers distributed by teachers or center staff. Potential participants were told that they would receive $10 US for participating and that they would also receive a raw listening score on a type of listening task that appears on the MET if they requested it.

All sessions were held in a computer lab, and prior to each session each computer had web browser pages set to a form of the listening test, a version of the forward digit span test, a version of the backward digit span test, and the website providing access to Shapebuilder and Blockspan. A participant card was placed at each computer station which included each participants' unique ID, as well as the URLs for each measure in case a web page was accidentally closed. All sessions were proctored solely by the author in accordance with University of Maryland IRB regulations about the need for proctors to have successfully completed the University of Maryland's IRB certification process. At the start of each session a brief explanation of the study was provided, and all participants were given an English and Spanish version of the consent form. After reading the consent form(s), if they agreed to participate, they were asked to sign the English version and keep the Spanish version for their records. All measures were administered in the same order in both locations 1) listening test (preceded by biographical background questions), 2) forward digit span, 3) backward digit span, 4) Blockspan, and 5) Shapebuilder. Participants were told to notify the proctor when they completed each task in order to be given a short explanation of each upcoming task, and in the case of Blockspan and Shapebuilder, to give the participants the Spanish-language translation of the directions. Upon completing all tasks, the participants were given $10 US and signed a payment register to indicate receipt of the payment.

It is not ideal for all measures to be administered in the same order, but the constraints of the data collection context (one proctor and up to 15 participants per session) did not allow for a counter-balanced order. However, only a handful of participants were unable to complete the tasks in the two allotted hours, and with the vast majority completing the tasks between one hour twenty minutes and one hour forty-five minutes, fatigue or wavering attention was not observed (and in hindsight, four or five listening test forms with 50 items each instead of seven test forms with 40 items each—or including 8–10 common link items—would have been feasible and would have increased the power for the analyses described below).

4.4 Analysis

Data analysis involved two phases. First, participant performance on the five instruments (TCI, Blockspan, Shapebuilder, backward auditory digit span, and forward auditory digit span) were scored and the reliability for each measure estimated.

The dichotomous multiple-choice item results were scored using the Rasch model with the Winsteps 3.93.2 software program to obtain item difficulty estimates and person ability estimates, which served as dependent variables for research questions focusing on working memory or predictor variables in other analyses. Additionally, a polytomous Rasch model (i.e., the partial credit model) was run to generate person ability scores on the constructed response items, which were the basis for the dependent variable for several research questions. All constructed-response items were scored on a 0–2 scale, with the decision being made to include a three-step scale based on responses in the pilot which showed that some participants are able to provide partial answers (e.g., "they will stay inside" vs. the expected "they will continue to *work* inside"), or in cases where the

correct answer was provided, but additional erroneous information was also included ("the man was late" vs. "the man was late because he missed the bus" in response to "what happened this morning?" when the actual reason for lateness was "his taxi got lost"). All CR item responses were scored by the author, but because there is an element of subjectivity in CR scoring, a subset of the CR responses was doubled scored: 600 by two other raters (with 400 of the items across additional raters overlapping, so a total of 800 responses total received two ratings). The absolute agreement and correlations among the raters were adequate (see section 5.1.1), although one limitation of the study was the inability of raters to discuss ratings when there were disagreements.

The working memory measures scores were generated by the website and reliability was estimated using the item-level responses provided to the author by the creator of the measures (reliability estimates are provided below in the results chapter). The FDS and BDS tasks were scored manually using Excel to match number strings provided by the participants to the keys. Partial credit was given when only one error was present in the string (regardless of it being a three or seven-element string). Participants' scores on the measures was the sum of points given for each item/element string.

After scoring was completed and reliability estimated, several types of inferential analyses were run, including a confirmatory factor analysis to look at the question of subskill separability and influence of format (section 5.3), a logistic regression to look at subskill ability's predictive role in CEFR level classification (section 5.5), and multiple regressions to look at the role of WM (section 5.6).

Chapter 5: Results

In this chapter, descriptive results and reliability estimates are provided for the five measures included in this study, with the sections that follow reporting on the results of the investigation of the four primary research questions being asked: 1) are implicature item keys accessible to target test-takers, 2) does a confirmatory factor analysis show evidence of separability of subskill between general listening and implicature listening, 3) do implicature items separate higher level language users better than general listening items, and 4) do we see evidence of a greater influence of working memory capacity for performance on implicature items than on general listening items.

5.1 Descriptive results and reliability estimates

Table 3 shows descriptive statistics for participant performance on the MC listening items, CR listening items, Shapebuilder, Blockspan, forward auditory digit span (FDS), and backward auditory digit span (BDS) tasks.

Table 3. Descriptive Statistics

Measure	N	Min	Max	Mean (Std. Error)	S.D.	Variance	Skewness (Std. error)	Kurtosis (std. error)
MC listening	251	2	19[1]	10.52 (.246)	3.894	15.163	.098 (.154)	-.907 (.306)
CR listening	251	2	39[2]	18.22 (.568)	9.001	81.014	.253 (.154)	-.942 (.306)
Shape-builder	248	66	2930	1362.32 (30.66)	482.927	233218.41	.228 (.155)	.024 (.308)
Block-Span	249	220	3000	1291.33 (31.094)	490.652	240739.16	.684 (.154)	.649 (.307)
FDS	254	0	20[1]	16.8 (.200)	3.184	10.137	-1.379 (.153)	2.958 (.304)
BDS	242	0	20[1]	12.00 (.416)	6.471	41.871	-.663 (.156)	-.794 (.312)

[1]Maximum possible score was 20; [2]Maximum possible score was 40

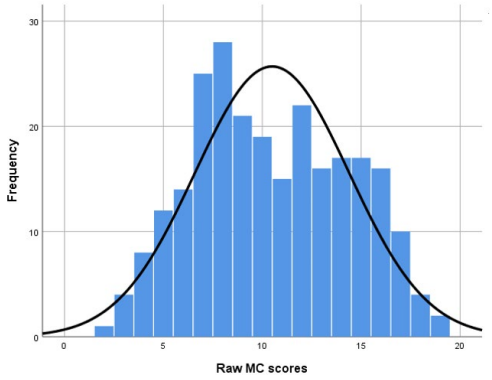
The distributions for MC and CR listening raw scores were adequate (see Figures 7 and 8 below for histograms of their distributions). There is no skewness for either when following the primary rule of thumb (Lomax, 2007) that skewness estimates between -1 and 1 are generally acceptable (with zero being the skewness estimate for a perfectly normal distribution), and the secondary rule of thumb that three times the standard error of the skewness estimate should be greater than the absolute value of the skewness estimate. For the MC values there were also no issues with kurtosis (the extent to which expected values in the tails of the distribution are present). For the CR values, there is some indication of kurtosis; the estimate is close to -1 and it does not meet the three times standard error rule of thumb (Lomax, 2007): i.e., |-.942| > .918, but the visualization of

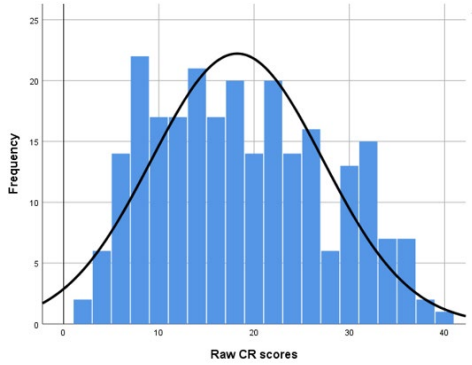
the data indicates that the degree of kurtosis exhibited in the data is not to the extent that it needs transformation.

Turning to the memory measures, assumptions of normality of the data distribution for the working memory measures appear to be met sufficiently (Figures 9 and 10). The distribution for Shapebuilder shows no signs of skewness or kurtosis. The distribution for Blockspan, while not showing kurtosis, does show some evidence of positive (right-side) skew. However, as with the CR item distribution, it does not necessitate transformation. The distributions for the forward auditory digit span and the backward auditory digit span task, however, were problematic to the degree of requiring adjustment and/or not including in the inferential analyses. For FDS, the skewness estimate (-1.379) falls into the category of being described as highly skewed, which is also readily apparent from the histogram (Figure 11). The frequency of perfect scores reflects the possiblity that some participants were recording the digits while they were being heard, rather than recalling at the end of the item—a possiblity that was not precluded either by the measure's online interface or by fact that pencils and paper (personal copy of consent form and/or identifcation card at their computer station) were available to the participants during the testing sessions.

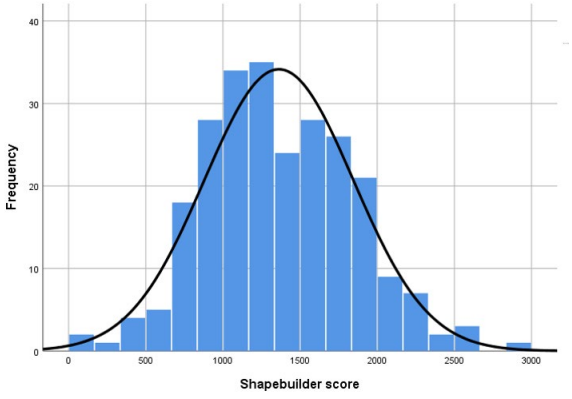
For the BDS task (Figure 12), the unexpectedly high rate of "zero" scores (i.e., responding incorrectly to all 20 items) appears to show that some participants—despite the direction being in English and Spanish—did not respond by providing digits in backward order but instead provided them in forward order (as in the FDS task, which immediately preceded the BDS task); this was confirmed by looking at the response data. This means that there can be little confidence in the scores given for these two STM

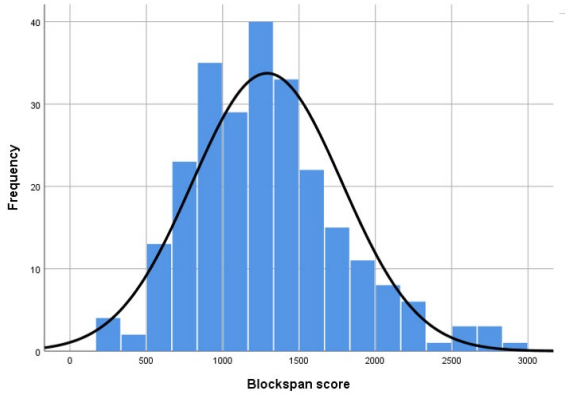
tasks; it is impossible to determine after the fact how many perfect scores on FDS are genuine versus how many are not. For this reason the FDS scores were not included in any of the analyses. The BDS scores are also problematic, but after discarding all the zero scores, they were included in some of the preliminary analyses (discussed in section 5.6 below), but again, those values must be considered questionable.



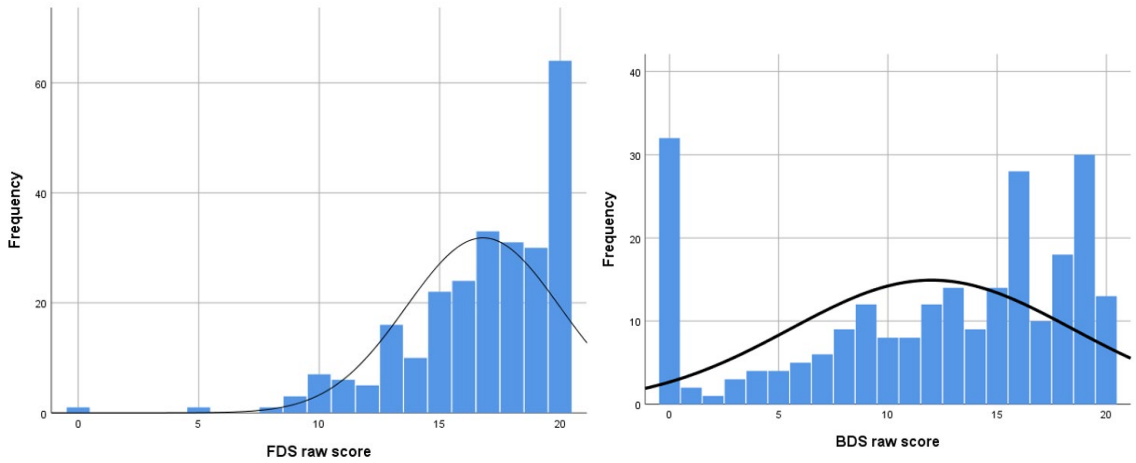


Figures 7 and 8. Distributions of MC Raw Scores and CR Raw Scores





Figures 9 and 10. Distributions of Shapebuilder Scores and Blockspan Scores



Figures 11 and 12. Distributions of FDS Raw Scores and BDS Raw Scores

The data were all checked for outliers and none were detected in the MC raw or CR raw distributions. In the working memory measures, one high-score outlier was detected in the Shapebuilder data and seven high-score outliers were detected in the Blockspan data, but these were flagged by the 1.5-standard deviation threshold, not by the 3.0-standard deviation threshold. Considering that, and that the overall distributions seen in Figures 9 and 10 above were acceptable, they were not deemed to be extreme outliers and were kept in the data. As mentioned already, the FDS scores were so problematic they are not being included in the analyses, so an outlier check was irrelevant. The BDS distribution, while also problematic, did not indicate outliers, which was not surprising, as the score range was restricted.

The reliability for the two working memory measures and BDS were estimated using Cronbach's alpha. With all measures reporting Cronbach alpha values of >.75, the reliability was deemed adequate.

Table 4. Working Memory and BDS Task Reliability Estimates

Measure	N	Number of items	Cronbach's alpha
Shapebuilder	221[1]	26	.836
Blockspan	217[2]	40	.780
BDS	205[3]	20	.857

[1]calculated from 221 participants who responded to all 26 items
[2]calculated from 217 participants who responded to all 40 items
[3]calculated after removing 34 participants who received total raw scores of zero

5.1.1 Constructed-response scoring

Before summarizing the Rasch analyses of the listening test, it is necessary to provide an overview of how the CR responses were scored. The 251 participants included in the listening test analyses had the opportunity to provide a total of 5,020 responses to the CR items. However, rather than ask participants to guess, they were given the opportunity to skip CR items (and MC items) if they did not have an answer. This ability to skip items resulted in 492 unanswered CR items, which means there were 4,528 CR responses to be scored. The items were all scored initially by the author on a 0–2 scale: zero points were given for answers that were completely wrong, one point was given for answers that were partially correct (or were correct but also included incorrect information) and two points were given for fully correct responses. The starting point for "fully correct" responses was the key from the MC format version, but as equally correct synonyms or wordings were encountered in the participant responses, they were added to a CR scoring guide document that was provided to additional raters. The scoring guide also included two overarching rules for partial scores:

> 1. Providing part of the answer but not enough to be fully correct, e.g., writing in response to a conversation about the high cost of living in the city (item LGC_33) that "they are discussing their opinions of the city" or "they are discussing the price of gas" when the full credit key is "high prices in the city" or "the cost of living in the city."

2. Providing the correct answer but going beyond that to introduce an element that was not correct, e.g., writing a response to a conversation about why a man was late to work (item LGC_52) that says "the regular bus didn't show up and he took a taxi and it got a flat tire and it lost the way" when the taxi part is correct but a bus not showing up was not part of the conversation.

As examples of acceptable partial credit responses were encountered, they were added to the scoring guide. The scoring guide also emphasized that spelling and grammar errors were not constraints on achieving full or partial credit, although there were a small number of responses where incorrect tense proved problematic and was counted against the participant, e.g., when something happening in the future or past was central to the point being tested.

After the responses were given their initial scores by the author, a subset of 600 responses were sent to a second and a third rater to verify the reliability of the scores, with 400 of the 600 responses being sent to both raters. This means that 800 responses received at least two ratings (roughly 17% of the total dataset) and 400 responses received at least three ratings. Both secondary raters were selected on the basis of their extensive experience as item writers and as ESL/EFL writing raters. The responses that were sent to the additional raters were semi-randomly selected, with one factor for selection being to send responses from participants who did not have missing data, in order to maximize the amount of overlap.

The raters were provided a spreadsheet in which to enter their responses (without seeing what the first rating was), along with the item content and the scoring guide with its explanation of scoring and examples (as summarized above). The double raters were also provided the initial ratings given by the author on a separate sheet, with a direction to not look at those ratings until after giving their own ratings. A limitation to this

double-rating process was that logistic constraints did not allow for norming sessions, but both raters were told to contact the author with questions if they encountered problematic responses for which they felt there was insufficient or unclear guidance in the CR scoring document. A small number of emails were exchanged along those lines, but there was not a formal training or norming discussion step. However, as Table 5 shows, the degree of reliability exhibited by the three raters as seen through absolute agreement (70.7–76.2%) and Pearson correlations (.773–.795) is high enough to be able to say that there was reasonable reliability for the human scoring of the CR responses. Although it is inevitable that there is some degree of error in the CR ratings, given a three-point scale, the adjacent agreement values should be expected to be at ceiling. The lack of adjacent agreement that was found in some cases revealed oversights in scoring by one rater; therefore, it is possible that in the single-scored data, a small percentage will contain non-adjacent scoring errors.

Table 5. Summary of Double-rating of CR Responses

Rater Pairs	Number of shared ratings	Absolute agreement %	Adjacent agreement*	Correlation (Pearson)
R1:R2	600	76.2	95.2	0.789
R1:R3	600	70.7	95.4	0.773
R2:R3	400	74.5	95.0	0.795

As an additional comment on the CR rating, an indication that a 0–2 rating scale would need to be investigated further for a context where the stakes are higher was the lower rate of use of the partial credit value. The reliability and separation values obtained from the Rasch analyses are slightly better for the CR items than the MC items (Rasch results discussed in detail in section 5.1.2 below), which would mean that the CR item versions contribute to discriminating between levels of learners. But in terms of a traditional scale

diagnostic (Figure 13 below), we do not see the information curves for each score point performing distinctly from each other. Therefore, the question of the value of including a partial credit category is a natural one. While partial credit was awarded on the CR items at a lower rate (about 16%) than zero credit (about 41%) or full credit (about 42%)—and is certainly lower than in essays or more extended writing, where demonstrating partial ability is expected of many, if not most, test-takers—16% does not seem to fit in a category of "rarely." It appears that further investigation beyond this study is needed to ascertain whether dichotomous scoring is more justifiable.

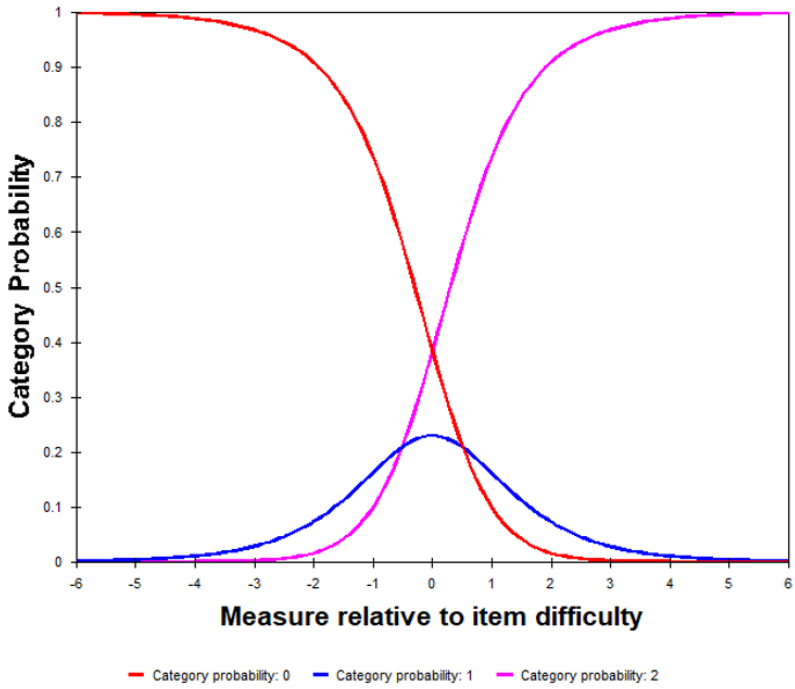


Figure 13. CR Scale Category Probability

5.1.2 Rasch results

Once the CR items were scored and a subset were double-rated with an acceptable degree of reliability, Rasch analyses were conducted on the listening test items using the Winsteps program, version 3.93.2. Analyses were conducted separately on the MC items as a group of 60 items and the CR items as a group of 60 items. Ability estimates were

also generated for MC and CR items together—although conflation is a possible issue in those results (i.e., the CR and MC items based on the same stimulus were not wholly independent), so the results of the full dataset analyses were not the basis for the item difficulty and person ability estimates that are being discussed in this study. The results from the separate 60-item MC analyses and 60-item CR analyses were (which are reported in Appendices E and F).

One study design flaw emerged while doing the Rasch analyses. Despite attempts to create full connectivity across the 120 items via the seven 40-item forms, the initial Rasch output indicated there were two subsets in the data. Following (Linacre, 2017, p. 101), this was resolved by including two dummy lines of data for MC where correct and incorrect responses alternated (i.e., 10101…/01010…) and three dummy lines of data for the partial credit analyses (i.e., 012012…/120120…/201201…). This ensured connectivity and altered the Rasch estimates in an extremely minimal way. With the dummy participant lines in the data set to ensure the absence of subset results, the Rasch analyses showed that the items in both MC and CR formats performed adequately, as is seen in Table 6.

Table 6. Item Difficulty Measures for TCI by Item Format

Item Format	**Item Separation**	**Item Reliability**	**Item Difficulty S. E.**	**Person Separation**	**Person reliability**
CR (k=60)	4.03	0.94	0.17	2.07	0.81
MC (k=60)	3.64	0.93	0.27	1.59	0.72

The data were examined for misfitting items—items whose estimates contribute minimally or detract from the overall model—by looking at the point-measure correlation and outfit mean square statistic (see appendices E and F for the full MC and CR item

measure tables). Regarding the choice of using the outfit rather than the infit statistic (Rasch provides both for all items and people), whereas the infit statistic is adjusted to give more weight to unexpected responses on items that closely match an item's estimated difficulty, the outfit statistic is not weighted in this way. This means outfit is more sensitive to unexpected responses on very easy or very difficult items and is the fit statistic that Linacre (2017, p. 554) recommends using (unless special circumstances require preference for the infit measure).

Only one (LIC_55) of the sixty items in the CR data had a point measure correlation value below 0.20 (the industry norm for a minimum). Four of the sixty MC items had point measure values below 0.20 (LIM_15, LIM_25, LGM_28, and LIM_56). However, because these values were not negative (i.e., detracting from the model) and because of the investigative nature of this study, they were not removed from the data as they would normally be in a higher-stakes test development situation where the highest precision of the measure is being sought. Similarly, when examining the outfit mean-square statistic, the more lenient acceptable value range of 0.5–1.5 (Wright & Linacre, 1994) was used, rather than the 0.7–1.20 value range recommended for higher stakes situations (Bond & Box, 2001; McNamara, 1996). Using the more lenient 0.5–1.5 range for this experimental context, only four items in the MC data (same as those flagged for low point-measure correlation: LIM_15, LGM_25, LGM_28, and LIM_56) and only six items in the CR data were flagged (LIC_01, LIC_05, LIC_14, LGC_37, LGC_46, and LIC_55). However, only one of the items' outfit estimate was greater than 2.0 (LIC_55), the point at which item performance is not only not contributing to the over model

estimates, but is detracting from them, so as with point-measure correlation, the decision was made to keep them in the analysis.

When looking at participants, the Rasch analyses indicated a small number of participants (i.e., 1033 and 1252 in the CR data and 1043 and 1064 in the MC data) who did not appear to fit the model well (negative point-measure correlation and outfit meansquare values > 2.0). However, given the experimental nature of the study, and following Linacre (2019, p. 525), whose advice on iteratively removing data is to do so only if the data is "really, really bad" (verified by removing the misfits, running the analysis again and checking for noticeable change). The two worst fitting participants from the MC data and the CR data were removed, the analyses were re-run, and when it was verified that standard error, separation, and reliability estimates did not change, and a check of individual item difficulty and person ability measures showed no change, or changes were of the nature of increases or decreases of about 1/100 of a logit, it appeared that those misfits were not actually detracting from the overall model and were left in the data. Figure 14 shows histograms of person ability over CR item difficulty (left) and MC item difficulty (right).

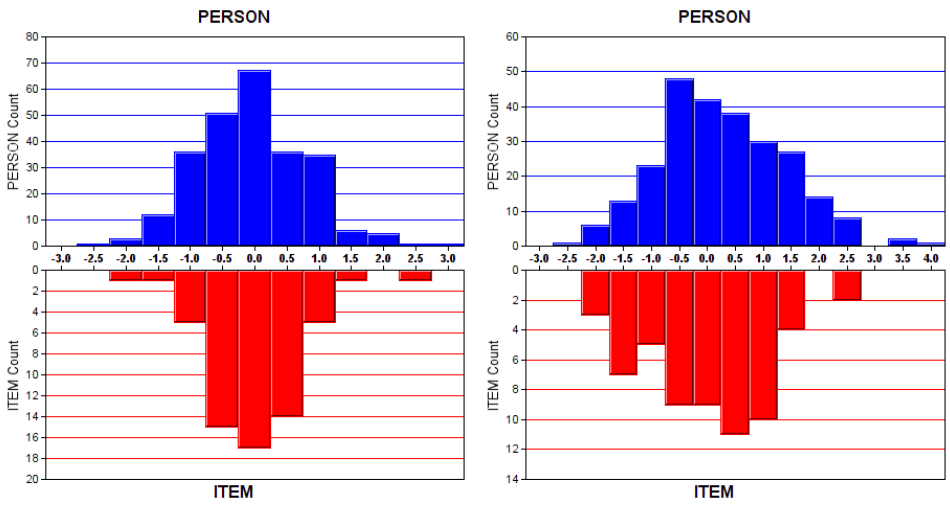


Figure 14. Histograms of Person Ability over CR and MC Item Difficulty

In sum, the Rasch analyses show that the listening test items, in both multiple-choice and constructed-response versions, worked sufficiently well to obtain a listening ability estimate for the participants. There were more misfitting participants than items, but this is to be expected in any testing situation, and in order to answer the questions being asked, in consideration of verifying the lack of harm they did to the Rasch estimates, they were retained in the data.

5.2 Dimensionality of the listening test

Because a core question in this study is the separability of implicature items from non-implicature items, additional analyses were run using the Winsteps 3.93.2 program to look at the possibility of multidimensionality in the listening test. Multidimensionality can be detected in Rasch—the model of which assumes unidimensionality—by examining the residuals (the differences between observed data and expected values by the model) for items that are not part of the expected Rasch dimension (Linacre, 2018). In the case of the listening test used in this study, if multidimensionality were observed in accord with the implicature subskill-focused items versus the general subskill-focused items, that would be evidence for the two subskills. It is important to note that this

analysis is not akin to traditional factor analysis; Linacre describes it as an examination of the contrasts between two factors, not an examination of the loading of items on factors. Furthermore, this analysis is recommended as a better way of looking at multidimensionality in Rasch results than solely looking at the fit of individual items. Misfit statistics that cluster on particular items might be an indicator of a commonality between those items, but as fit statistics are related to individual items, it is too subtle a method for making inferences about dimensionality (Linacre, 2018). Rather, it is recommended that the standardized residuals are examined with a principal components analysis to determine if unexplained variance is clustered on a meaningful number of items (i.e., if more than 3 or 4 items are explaining variance together, it may be due to multidimensionality rather than just error).

The analysis was run on both the multiple-choice items and constructed-response items. Table 7 shows the results for the multiple-choice items.

Table 7. Principal Components Analysis of Residual Variance for MC Items

	Eigenvalue*	**Observed**	**Expected**
Total raw variance in observations	86.2	100.0%	100.0%
Raw variance explained by measures	26.2	30.4%	30.2%
Raw variance explained by persons	10.9	12.6%	12.5%
Raw variance explained by items	15.3	17.8%	17.7%
Raw unexplained variance (total)	60.0	69.6%	69.8%
Unexplained variance in 1st contrast	3.3	3.8%	5.4%
2nd contrast	2.3	2.7%	3.9%
3rd contrast	2.0	2.3%	3.4%
4th contrast	1.9	2.3%	3.2%
5th contrast	1.9	2.2%	3.1%

*Eigenvalue = item information

The first values that need to be examined in this analysis's output is whether the observed and expected values (two columns on the right) for the unshaded top rows are equal or

are very similar, which they are for the multiple-choice items. This match in values tells us that the amount of explained variance is correct.

The first shaded row, "Raw unexplained variance" and the ensuing shaded "contrast" rows are what are of most interest. Per Linacre, the amount of variance explained by the Rasch model is not relevant here—it is what *other* dimensions are accounting for—what is unexplained and yet is correlating—that we want to examine. When interpreting these values, however, it should be noted that the value for raw unexplained variance total is always equal to the number of items in the measurement tool—60 in this case, because each item contributes one "unit" of variance; the computation is set up in this way as a matter of convenience (Linacre, 2018).

Each unexplained variance and contrast row identifies correlated "unexplained" activity among contrasting clusters of items; that is, items whose residuals correlate with each other. These clusters could be solely accidental or they could be a secondary dimension. The rule of thumb that Linacre proposes is that the secondary dimension must have the strength of at least two items—if the Eigenvalue is below 2.0, it is a contrast caused by one idiosyncratic item. If the data fits the Rasch model perfectly, every component in these analyses would be one item strong. That is not the case in this table, as we see that the first component (or contrast) is 3.3 items strong, and two others are about 2.0 items strong. But taking Linacre's rule of thumb that these values need to be greater than two or three to be worthy of serious examination of a secondary dimension, there does not appear to be evidence of a secondary dimension. This lack of evidence is reinforced when considering that the listening test consisted of 30 implicature items and 30 general items, and among the implicature items, there was an attempt to vary them by

"degree of implicature"—but we do not see items correlated together by degree or even subskill. When the three items with the strongest loadings in the first contrast/component were investigated, they were two general items and one implicature (LIM_19, LGM_51, and LGM_29), further indicating a lack of dimensionality by subskill.

The same analysis was run on the CR items; Table 8 below shows the output.

Table 8. Principal Components Analysis of Residual Variance for CR Items

	Eigenvalue*	**Observed**	**Expected**
Total raw variance in observations	98.0	100.0%	100.0%
Raw variance explained by measures	38.0	38.8%	38.8%
Raw variance explained by persons	26.2	26.7%	26.7%
Raw variance explained by items	11.9	12.1%	12.1%
Raw unexplained variance (total)	60.0	61.2%	61.2%
Unexplained variance in 1st contrast	3.1	3.2%	5.2%
2nd contrast	2.7	2.7%	4.5%
3rd contrast	2.4	2.4%	4.0%
4th contrast	2.1	2.1%	3.5%
5th contrast	2.0	2.1%	3.4%

*Eigenvalue = item information

As with the MC items, the observed and expected values in the unshaded rows are equal. Looking at the eigenvalues for the first five contrasts, while there is slightly more clustering than with the MC values, we again see values in the 2–3 range, which indicates a very low likelihood of a dimension attributable to implicature that is creating a cluster. In fact, when the three items clustering on the first contrast were examined, as with the MC output, they consisted of both implicature (LIC_56 and LIC_20) and general items (LGC_53).

The question of dimensionality will be addressed again below in the section detailing the CFA results, but what this Rasch output indicates is that there does not appear to be a dimension attributable to implicature evident in the listening test.

5.3 Performance on implicature items

Answering the first research question, which asked whether the keys for multiple-choice items assessing implicature are accessible to the target test-takers, required an examination of how items performed by format and how target test-takers responded to the implicature items in the CR format.

5.3.1 Comparing CR and MC performance

Before examining item-level CR performance, a comparison of CR- and MC-item performance will help set the context. Figure 15 shows a scatterplot of the Rasch item difficulty estimates for the 60 items in the TCI with CR difficulty on the y-axis and MC difficulty on the x-axis. What the scatterplot shows is that when items tended to be difficult in MC format, they also tended to be difficult in the CR format, and vice-versa. This is relevant to the question of whether CR item keys are accessible to test-takers for implicature items in particular, because if those keys are not accessible, the CR items should trend more difficult and only a weak correlation would be observed.

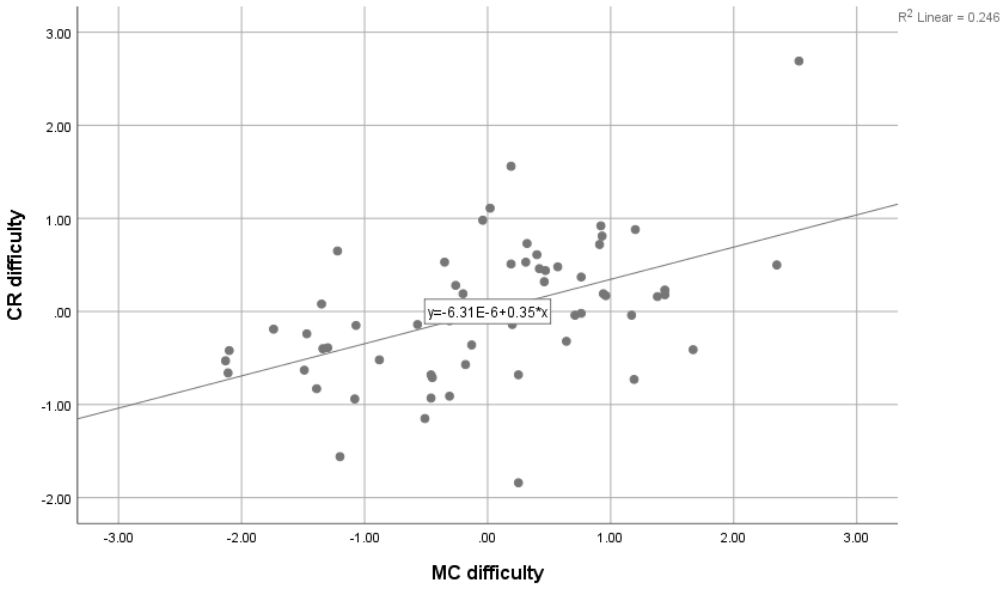


Figure 15. MC and CR Item Rasch Difficulty Estimates

The difference between overall item performance in CR and MC formats was, as predicted, minimal. There were idiosyncratic differences at the item level and the moderate Pearson correlation of .496 (significant at < .01) was slightly lower than expected—but the overall differences were small, and the lack of meaningful differences in means was confirmed via paired samples T tests that had statistically non-significant results, which are shown in Table 9.

Table 9. Summary of Paired Sample T Tests

Items	**MC difficulty estimate mean**	**CR difficulty estimate mean**	**Correlation of MC to CR difficulty**	**Sig of difference of means**
60 (implicature & general)	0.0005	0.0002	0.496**	0.998
30 (implicature only)	0.1837	0.3550	0.620**	0.291
30 (general only)	-0.1827	-0.3547	0.315 (not significant)	0.367

The minimal effect of format is also seen in the box plot in Figure 16 below, which displays the items by format and subskill. We see that implicature items (in blue on the left of each format pairing) were slightly more difficult than general items in both formats, but the difficulty across formats by subskill is minimal.

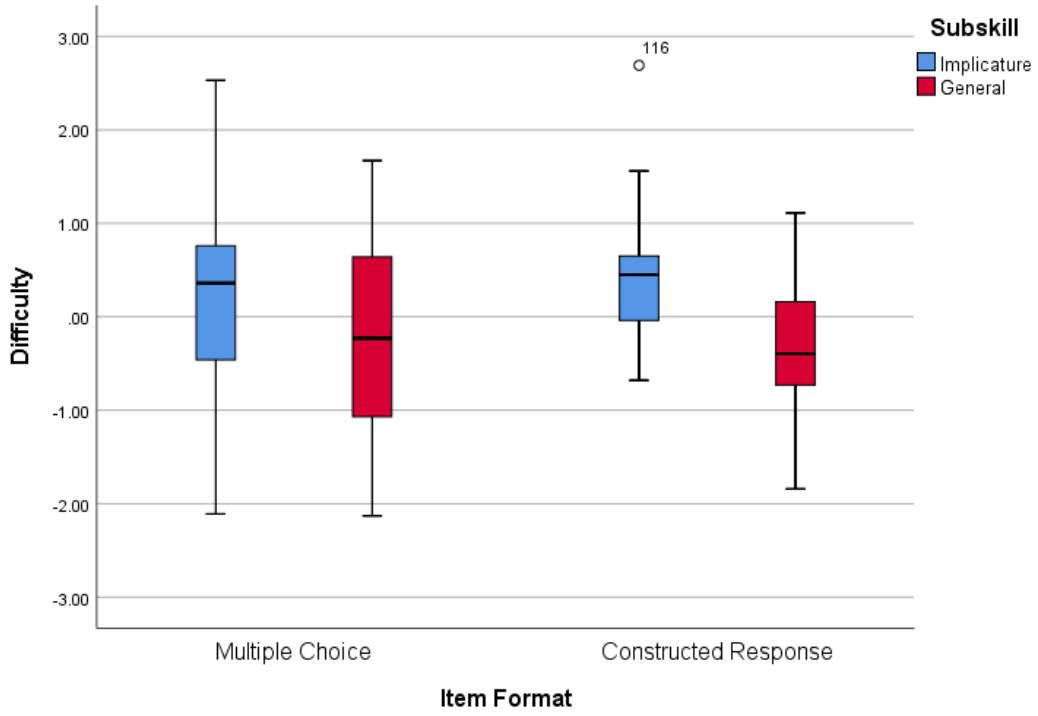


Figure 16. Item Difficulties by Format and Subskill

This expected lack of added difficulty resulting from CR format is potentially explained by a rigorous method of multiple-choice item development. With at least three experienced language testers independently reviewing items, whose method of review involves engaging with the items in an auditory mode first, and who as part of the review process are asked to generate their tip-of-tongue responses to the stem, the likelihood of inaccessible keys for test takers (TTs) is diminished—and further diminished by piloting processes.

However, it is of interest to look more granularly at the accessibility question. A frequent criticism of MC items is that TTs are provided the answer among an array of options, and it is unknown whether they would be able to generate this same answer if left to their own devices; as mentioned above, this topic has received surprisingly little attention in the listening assessment literature (Buck, 1991; Wu, 1998). To address this

question, the highest performing subset of TTs were examined. The rationale for choosing the highest performers may seem like biasing the outcome, but the items were developed for high intermediate to low advanced listeners (i.e., CEFR B2 and C1), so an examination of lower intermediate level responses would not be informative.

Despite efforts to recruit a sample of TTs that was split evenly by thirds across the CEFR B1, B2 and C1 levels, the majority of participants were in the B1 to low B2 category. Therefore, the method chosen for looking at individual responses was the top 15 TTs (top 6%) plus the 13 participants who were independently identified (by their MET scores) as being C1. This resulted in 24 sets of individual responses in the subset, because four of the C1 MET participants were also in the top 15 overall. Raw performance of these higher ability 24 individual (9.6% of the total sample) on their CR implicature items is summarized in Table 10 below, noting that because the listening test was limited to 40 items, half CR/half MC, and half implicature/half general, each participant was only exposed to 10 implicature items in CR format. This means that there is an uneven number of top 24 respondents on the 30 implicature items in the CR format (from 4 to 11, as is shown in the "n" column).

Table 10. Raw Score Performance by Top Performers on CR Implicature Items

Item ID	Imp. Degree	Imp. Type	N	% TTs receiving full credit	% TTs receiving zero credit	Rasch item difficulty (CR format)	Rasch item difficulty (MC format)	CR:MC*
LIC_03	High	Relevance	10	70.0%	30.0%	0.5	2.35	-1.85
LIC_06	High	Relevance	11	81.8%	0.0%	-0.04	0.18	-0.22
LIC_08	High	Irony	11	81.8%	9.1%	0.88	1.2	-0.32
LIC_15	High	indirect criticism	6	33.3%	0.0%	0.44	0.47	-0.03
LIC_18	High	Relevance (avoidance)	6	50.0%	0.0%	-0.24	-1.47	1.23

LIC_22	High	indirect criticism	7	42.9%	57.1%	1.56	0.19	1.37
LIC_24	High	Relevance	7	28.6%	57.1%	0.81	0.93	-0.12
LIC_56	High	Relevance	7	14.3%	85.7%	2.69	2.53	0.16
LIC_02	Med	Relevance	11	100.0%	0.0%	-0.63	-1.49	0.86
LIC_05	Med	indirect criticism	11	72.7%	27.3%	0.46	0.42	0.04
LIC_09	Med	Relevance	11	81.8%	18.2%	0.53	0.31	0.22
LIC_10	Med	Irony	10	90.0%	0.0%	-0.04	0.71	-0.75
LIC_12	Med	indirect criticism	6	66.6%	33.3%	0.18	1.44	-1.26
LIC_13	Med	indirect criticism	6	66.6%	33.3%	0.53	-0.35	0.88
LIC_14	Med	Irony	5	80.0%	20.0%	-0.14	-0.57	0.43
LIC_19	Med	indirect criticism	6	83.3%	16.6%	-0.66	-2.11	1.45
LIC_23	Med	Irony	7	57.1%	0.0%	-0.02	0.76	-0.78
LIC_55	Med	Relevance	7	85.7%	14.3%	-0.68	0.25	-0.93
LIC_59	Med	indirect denial	7	71.4%	28.6%	0.92	0.92	0
LIC_60	Med	indirect criticism	7	57.1%	14.3%	0.65	-1.22	1.87
LIC_01	Low	Relevance	11	63.6%	0.0%	-0.68	-0.46	-0.22
LIC_04	Low	indirect criticism	11	81.8%	0.0%	0.08	-1.35	1.43
LIC_07	Low	indirect criticism	11	90.9%	9.1%	0.51	0.19	0.32
LIC_11	Low	indirect criticism	6	66.6%	33.3%	0.48	0.57	-0.09
LIC_16	Low	indirect criticism	6	83.3%	16.6%	0.61	0.4	0.21
LIC_17	Low	Relevance	5	80.0%	20.0%	0.37	0.76	-0.39
LIC_20	Low	Irony	4	50.0%	25.0%	0.72	0.91	-0.19
LIC_21	Low	Relevance	7	57.1%	0.0%	-0.19	-1.74	1.55
LIC_57	Low	indirect criticism	7	42.9%	14.3%	0.32	0.46	-0.14
LIC_58	Low	indirect criticism	5	80.0%	20.0%	0.73	0.32	0.41

*negative values indicate less difficult in CR; positive values indicate more difficult in CR (also note that comparison of Rasch estimate format differences are from the two separate format analyses, so should be interpreted with caution)

The raw performances on implicature items in CR format summarized in the table above shows that only one of the thirty items testing an implicature chosen by the test

developers was not accessible to the participants. Examination of the item, LIC_56—which was part of the round 2 development that was not piloted and is highlighted in grey in the table above—shows an item that was problematic. The world knowledge/background experience necessary for understanding this item in real time (borrowing a snow shovel) was clearly inaccessible to these participants in South America. This serves as a reminder of something that standardized test developers are already aware of when testing implicature: there is a danger of crossing into world experience that makes the item inaccessible even to high level language learners who likely have little problem processing all of the language in the utterance. However, only one flawed/inaccessible item out of thirty does seem to show that the tested implicatures in these items were accessible. It also seems to support the notion of test developers' ability to create listening test material where coverage of "inferences that are unambiguously present," while also countering the notion that the nature of inferences is that all are free to infer differently; this of course leads to the discussion of the nature of the inference—and this will be touched on below. But if one makes the uncontroversial assertion that conversational implicature is part of daily communication, it is worthwhile to demonstrate this accessibility and convergence of interpretations for items used in a scripted listening proficiency test.

What is also instructive from Table 10 is to look at the final two columns—MC Rasch difficulty estimates and the difference between CR and MC. While it was stated above that overall performance differences in difficulty were minimal between MC and CR, the differences in format for implicature items was greater than for general items, and there are clearly idiosyncratic differences at the item level. Thirteen of the thirty

implicature items had differences in difficulty that were greater than half a logit, with most of them being more difficult in the CR than MC version. A closer investigation of responses is merited and the performance of items LIC_04, LIC_12, LIC_18, and LIC_21 will be described in detail in the discussion section below.

An examination of the aggregate Rasch results by subskill in Table 11 below shows similar performance as well.

Table 11. Aggregate Rasch Results by Item Subskill

Item Type	**Item Separation**	**Item Reliability**	**Item Difficulty S. E.**	**Person Separation**	**Person reliability**	**Person Ability S. E.**
General (k=30)	3.65	0.93	0.22	1.95	0.79	0.44
Implicature (k=30)	3.67	0.93	0.22	1.61	0.72	0.41

The items, while showing some differential effect in the person statistics (slightly better person reliability and ability to separate the test-takers into distinct groups for the general items), perform quite similarly in aggregate. There are differences within the implicature items' content and context, either by types of implicature or degree, but what seems to be clear is that there are no discernible patterns when language processing load is kept consistent. That is, varying by "degree" or "type" does not seem to add to or detract from difficulty within an assortment of conversational implicature items. Again, type of inference and degree of inference were not the targets of any of the research questions, but in development of the items, type in particular could not be ignored—there is a wide range of inference types and some attention had to be paid to their distribution. What was done differently in this study is that, following suggestions in Kasper (1984), a degree of inference evaluation was made, and performance by degree was examined even

though it was not the focus of a research question. There was only the most minimal of trends in mean difference by degree. It therefore appears that reliably predicting difficulty of inference items by "degree" is likely to be of only moderate usefulness (Figures 17 and 18 below).

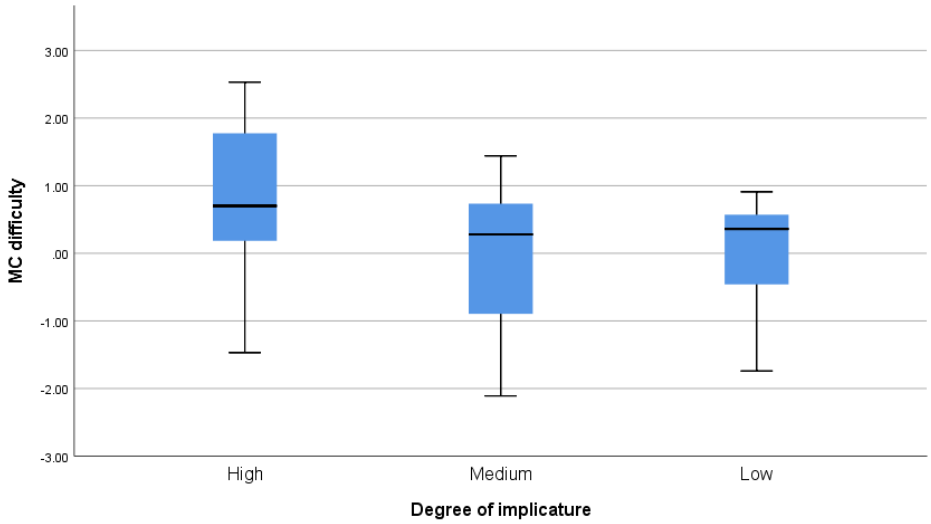


Figure 17. Rasch Difficulty by Degree of Implicature for MC Implicature Items

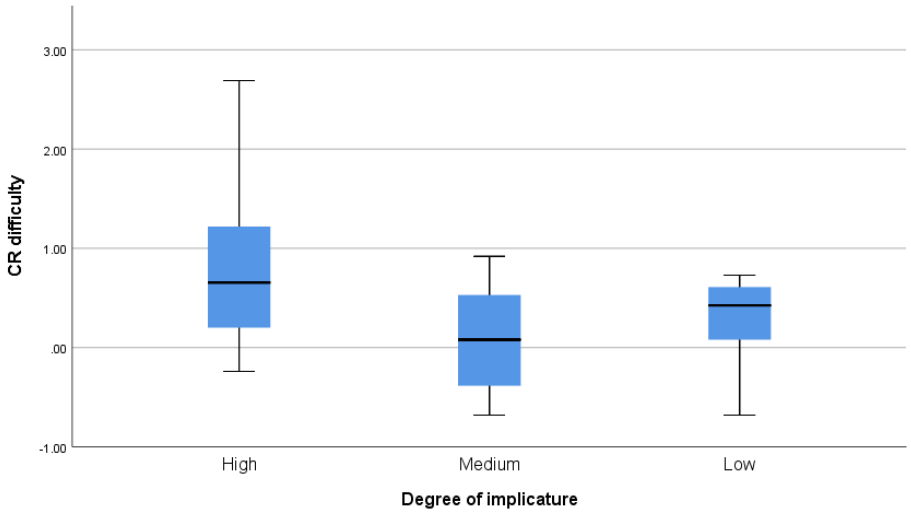


Figure 18. Rasch Difficulty by Degree of Implicature for CR Implicature Items

5.3.2 Logit models on implicature items

In order to look at the question of the possible existence of an implicature construct more granularly, i.e., at a level below the aggregate Rasch analyses and the descriptive analysis

of CR responses from high-level performers described above, a series of logistic regressions were run on the dependent dichotomous correct/incorrect response for each multiple-choice implicature item. The purpose of these analyses was to examine whether participants' performance on the other MC implicature items in their assigned version of the test was able to predict whether they would respond correctly to each individual implicature item.

Before running these logistic regressions, two global scores based on raw scores were generated per participant: their total multiple-choice implicature correct and their total multiple-choice general correct. The total multiple-choice implicature correct for a particular implicature item then had that result removed. That is, the implicature global score was defined by the sum of implicature correct responses minus the contribution of each distinct implicature item, and the general global score was a sum of all correct responses on the general MC items. The equations for the logistic models using these global scores can be represented as the equation in (1), where item1 = Implicature item 01, item2 = implicature item 02, and so on, for all thirty implicature items:

$$Y_{item1} = 1 = B_0 + B(\text{sumI-item1}) + B(\text{sumG}) \quad (1)$$

These analyses tested whether the probability of a correct answer on each implicature MC response was more predictable from other implicature correct or more predictable from the general MC items sum. If the former, there is evidence of an implicature skill; if the latter, evidence is provided that the probability of understanding an implicature is influenced more by having higher general proficiency than a separate identifiable implicature skill. The results point more towards the latter than the former, as is summarized in the tables below. Although it must be noted that because there were

seven test forms used in this study, with only ten multiple-choice implicature and ten multiple-choice general items on each, the analyses were not run on a uniform set of data. However, the trend does appear across all three sets of analyses and appears meaningful: from the thirty logistic regressions that were run, only in six instances was predictability of a correct answer improved by looking at performance on other implicature items as opposed to fifteen instances where predictability of a correct answer on the implicature items was improved by looking at general item performance (with predictability here being based on statistical significance of the "Implicature correct" variable and the "General correct" variable). A summary of the thirty analyses is presented below, broken across three tables (Tables 12–14) to match the grouping of participants who took the same MC implicature items and the same MC general items.

Table 12. Logistic Regressions on LIM_01–10

Item ID	**Rasch difficulty estimate**	**N (missing)**	**Sig of Imp variable**	**Exp(B) Imp variable**	**Sig of Gen variable**	**Exp(B) Gen variable**
LIM_01	-0.46	69 (1)	0.55	1.13	0.01**	1.50
LIM_02	-1.49	70 (0)	0.89	1.03	0.07*	1.43
LIM_03	2.35	65 (5)	0.53	0.85	0.13	1.38
LIM_04	-1.35	67 (3)	0.45	1.28	0.01**	2.03
LIM_05	0.42	68 (2)	0.39	1.17	0.21	1.19
LIM_06	0.18	69 (1)	0.29	1.21	0.80	1.04
LIM_07	0.19	69 (1)	0.13	1.30	0.16	1.22
LIM_08	1.2	68 (2)	0.49	0.88	0.02**	1.47
LIM_09	0.31	68 (2)	0.66	0.92	0.05*	1.33
LIM_10	0.64	70 (0)	0.24	1.24	0.81	0.97

*p ≤ .10, ** p≤ .05

For the first ten implicature items, which appeared on TCI forms 3 and 6, the logistic regressions showed no predictive ability of the implicature variable on whether or not the implicature items were answered correctly. But for five of the ten items (LIM_01,

LIM_02, LIM_04, LIM_08 and LIM_09), performance on general items, as captured by the General Variable, was predictive of correct answers.

Table 13. Logistic Regressions on LIM_11–20

Item ID	Rasch difficulty estimate	N (missing)	Sig of Imp variable	Exp(B) Imp variable	Sig of Gen variable	Exp(B) Gen variable
LIM_11	0.47	69 (3)	0.26	1.23	0.21	1.20
LIM_12	1.44	66 (6)	0.65	0.92	0.08*	1.34
LIM_13	-0.35	71 (1)	0.21	1.25	0.10*	1.27
LIM_14	-0.57	70 (2)	0.50	0.90	0.19	1.21
LIM_15	0.47	66 (6)	0.02**	1.66	0.01**	0.62
LIM_16	0.4	71 (1)	0.08*	1.41	0.00**	1.71
LIM_17	0.76	70 (2)	0.36	1.17	0.57	1.09
LIM_18	-1.47	69 (3)	0.11	1.52	0.03**	1.58
LIM_19	-2.11	72 (0)	0.26	0.75	0.00**	2.25
LIM_20	0.91	66 (6)	0.63	1.09	0.42	1.12

*p ≤ .10, ** p≤ .05

For implicature items 11–20, which appeared on TCI forms 1 and 4, the logistic regression results (Table 13 above) showed some predictive ability of the implicature variable on whether or not the implicature items were answered correctly; there were significant results at p < .10 for items LIM_15 and LIM_16. However, for six of the ten items, including LIM_15 and LIM_16, the general ability variable was predictive of correct answers. And the two implicature items that were predicted by the implicature score (LIM_15 and LIM_16) were also predicted by the general score.

Table 14. Logistic Regressions on LIM_21–24 & LIM_55–60

Item ID	Rasch difficulty estimate	N (missing)	Sig of Imp variable	Exp(B) Imp variable	Sig of Gen variable	Exp(B) Gen variable
LIM_21	-1.74	106 (2)	0.01**	1.95	0.37	1.17
LIM_22	0.19	95 (13)	0.57	1.08	0.07*	1.25
LIM_23	0.76	102 (6)	0.02**	1.42	0.20	1.17
LIM_24	0.93	104 (4)	0.06*	1.33	0.01**	1.46
LIM_55	0.25	97 (11)	0.11	1.25	0.64	1.06
LIM_56	2.53	104 (4)	0.72	1.06	0.74	1.06
LIM_57	0.46	103 (5)	0.25	1.16	0.29	1.13
LIM_58	0.32	99 (9)	0.57	0.92	0.00**	1.80
LIM_59	0.92	101 (7)	0.17	1.21	0.78	0.97
LIM_60	-1.22	102 (6)	0.05*	1.46	0.03**	1.40

*p ≤ .10, ** p≤ .05

The logistic regression analyses on the final ten implicature items (Table 14 above), which appeared in MC format on forms 2, 5, and 7, showed slightly more balance between predictive role of general ability versus implicature ability on keying implicature items. In this set, four of the ten implicature items (LIM_21, LIM_23, LIM_24, and LIM_60) were predicted by the implicature variable and four of the ten (LIM_22, LIM_24, LIM_58, and LIM60) were predicted by the general ability variable. But when looking across the full set of 30 items, the contrast is clear: only six items (20%) were predicted by performance on the implicature variable (and only two, LIM_21 and LIM_23, were solely predicted by the implicature score), whereas fourteen, or almost half, were predicted by general ability. This would lend credence to the view that performing well on implicature items is more related to linguistic ability than to a potential implicature ability.

However, 20% is not marginal and the six items in question were reviewed closely to consider whether there is some aspect of the content (e.g., topic or context) or

implicature (i.e., type or degree) that might tie them together. Table 15 below summarizes the metadata for the six items.

Table 15. Metadata for Implicature Items Predictable by Implicature Raw Scores

Item	Imp type[1]	Imp degree	Domain[2]	Topic	MC Rasch dif-Faculty	Point biser-ial	Infit	Outfit	CR: MC diff
LIM_15	IC	High	Occ	data sent wrong	0.47	0.05	1.4	1.74	-0.03
LIM_16	IC	Low	Occ	slow IT team	0.40	0.61	0.80	0.73	0.21
LIM_21	R	Low	Pub	Food shop-ping	-1.74	0.44	0.91	0.62	1.55
LIM_23	I	Med	Edu	Attend-ing conf.	0.76	0.52	0.95	0.85	-0.78
LIM_24	R	High	Per	painting garage	0.93	0.58	0.86	0.77	-0.12
LIM_60	IC	Med	Per	Re-cycling	-1.22	0.51	0.87	0.64	1.87

[1]IC=indirect criticism, R=Relevance, I=irony

[2]Occ=Occupational, Pub=Public, Edu=Educational, Per=Personal

When looking at implicature type, of which there were only three categories represented on the test (indirect criticism, relevance, and irony), all three types are represented in this small sample. The three levels of implicature degree (low, medium, high) are also represented equally, so that aspect of the items does not seem to link the six. There is no commonality with domain either—all four categories (Public, Personal, Educational, and Occupational) are seen in the six items. Topic is almost necessarily item-specific, discussion of which will be returned to below.

The columns on the right of Table 15 present Rasch information about the six items. This quantitative data does not appear to reveal any trends. For the Rasch MC difficulty estimate, two of the six (LIM_21 and LIM_60) were quite easy in the MC

format, two were on the more difficult side (LIM_23 and LIM_24), and two were moderately difficult. The point biserial and fit statistics are not indicative of a pattern, either—one of the six items (LIM_15) is weak in those areas, but the other five are all acceptable.

The logistic regressions were all run on the MC items, but the final column of Table 15 includes the difference in difficulty estimates between these six items in MC format and in CR format (similarly reported in section 5.3.1 above), to look at whether a format effect would be detectable for them (i.e., the majority easier in MC or more difficult in MC than CR). The positive values in that column show that an item was more difficult in CR, and three of the six items being examined here fall on that side, and three of the items fall on side of having negative values, and thus were easier in CR. While two of the positive values (for LIM_21 and LIM_60) are considerable—i.e., they were much harder in the CR format—there still does not appear to be an identifiable trend that would point to format effect as a commonality across the items, as one was quite a bit easier in CR (LIM_23) and three showed little difference across the two formats (LIM_15, LIM_16, and LIM_24).

The final aspect of the items that was looked at was the content (the full transcript of each item is in Appendix A). The content of the six items does not appear to reveal any trend in terms of difficult vocabulary or syntactic structures, nor is there a pattern by register, i.e., level of formality or power differential between the two speakers. The only discernible pattern was that potentially the topics of five of the six items were not part of the life experience of the majority of participants in their early 20s—LIM_15 and LIM_16 are very work specific, and LIM_23 is not only about preparing for travel to an

academic conference, but a conference in another country. LIM_24 and LIM_60 are about very specific household and apartment topics, painting a garage and putting out recyclable material properly in an apartment complex. On the other hand, four of these five had adequate fit statistics, and the consideration of the life experience of test takers is a crucial step in test development and is one that happened during the development of this test as well. But the question of the accessibility of topics is an important one and will be addressed in Chapter 6 within a broader discussion of the value of including implicature items in listening comprehension tests.

5.4 Confirmatory factor analysis results

Confirmatory factor analysis (CFA) is a type of statistical analysis that is commonly used in many fields, including language testing (Bae & Bachman, 1998; Llosa, 2007; Vafaee, 2016, etc.) to test hypotheses about relationships between observed measures and unobserved variables, which are often called latent variables (Brown, 2006). The goal of testing hypotheses is the key difference between confirmatory factor analysis and an exploratory factor analysis. In the case of the latter, the goal is often to uncover potential relationships between observed measures and an unobserved ability or latent factor. In this study, CFA was selected because the goal was to test hypotheses about whether models that separated the single factor of listening into implicature and non-implicature factors fit the data better than a single factor model. If this were the case, this would provide evidence for the existence of comprehension of implicature as a separate skill from general listening. Additionally, the possible role of item format, or method, was looked at, by correlating the errors of indicators that contained items of the same format. The correlated uniqueness model (Kenny, 1979) was the model used to look at the effect of method by correlating each indicator variable based on whether it comprised multiple-

choice or constructed-response items. Per Brown (2006), the correlated uniqueness model usually has at least two traits, as in the case in this study, and at least three methods. However, Brown states that "a 2T [trait] X 2M [method] model can be fit to the data if the factor loadings of the indicators loading on the same trait factor are constrained to equality," (p. 220), which was the case with the model used in this study.

Regarding sample size, factor analysis is generally considered a type of analysis that requires a large sample (Pett, Lackey, & Sullivan 2003), with rules of thumb stating that a 10:1 ratio of cases to indicators is a minimal threshold and that 20:1 should be the goal. All the CFA models run for this study had 8 indicators, each consisting of performance on five items, so the overall sample size of 251 appears to be sufficient (> 30:1 participant-to-indicator ratio). However, as discussed in the section above in regard to the Rasch results, the design of this study had a limitation in that the need to administer 60 items in both MC and CR format (meaning 120 distinct items in total), required multiple forms. That is, the 251 participants enrolled in the study did not all take the same listening test, and the absence of link items across the seven listening test forms prevented the Winsteps program from being able to provide Rasch estimates for indicator items that were not directly administered to participants. As a result, while all participants had a score for all indicators, it is not the case that all the scores were generated by the same bundle of items. Table 16 provides an explanation of the composition of the eight indicator variables.

Table 16. Item Composition of Indicator Variables Used in CFA Models

Indicator	**Composition**
imc1	Sum of participants' raw score total on **first** five **multiple-choice implicature** items they were given
imc2	Sum of participants' raw score total on **second** five **multiple-choice implicature** items they were given
gmc1	Sum of participants' raw score total on **first** five **multiple-choice general** items they were given
gmc2	Sum of participants' raw score total on **second** five **multiple-choice general** items they were given
icr1	Sum of participants' raw score total on **first** five **constructed-response implicature** items they were given
icr2	Sum of participants' raw score total on **second** five **constructed-response implicature** items they were given
gcr1	Sum of participants' raw score total on **first** five **constructed-response general** items they were given
gcr2	Sum of participants' raw score total on **second** five **constructed-response general** items they were given

Using the StataIC 15.1 software program, four models were compared with these data: a one-factor uncorrelated-by-method model; a one-factor correlated-by-method model; a two-factor uncorrelated-by-method model; and the correlated-uniqueness model (i.e., a two-factor correlated-by-method model). The correlated uniqueness model was hypothesized to be the best fitting model based on the assumption that implicature listening is distinct from non-implicature listening and that the different item formats (MC and CR) would account for some of the error. All of the models were estimated with maximum likelihood and are displaying standardized coefficients for the loadings. All of the models had one indicator variable (typically the first of the four or eight indicator variables) constrained to the latent variable to allow the models to converge. Each model is represented graphically in Figures 19–22 below, with factor loadings and error estimates included.

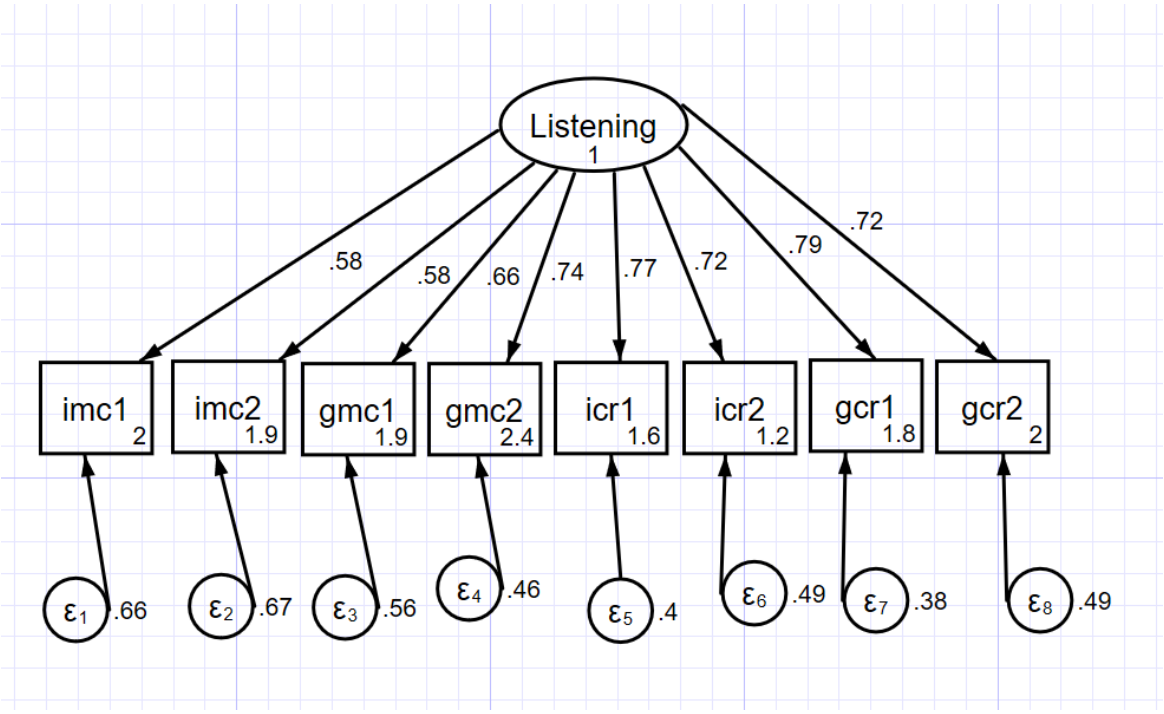


Figure 19. One Factor Model with Method Errors Uncorrelated

This first model shows fairly strong loadings from all indicators, ranging from .58 to .79, to the latent variable of listening, regardless of whether they are "i" implicature items or "g" general. The loadings from the second model, which was run with the method errors correlated, were slightly better (Figure 20 below). They ranged from .58 to .81 and in general were close to an average of .70 compared to an average of around .60 for the first model. These results indicate that the items in the listening test all seem to be testing at a broad level, and as designed, the latent trait of listening comprehension ability, and that when different item formats are correlated, the amount of error in the model is reduced. Goodness of fit indices were examined for the single factor models shown in Figures 19 and 20. The goodness of fit indices' values that were examined were Chi-square, root mean square error of approximation (RMSEA), the comparative fit index (CFI), Tucker-Lewis index (TLI), and Akaike's information criterion, and the Bayesian information

criterion (BIC). The values from the single factor models are discussed comparatively below in Table 19, to contrast the single factor and double factor models' fit simultaneously.

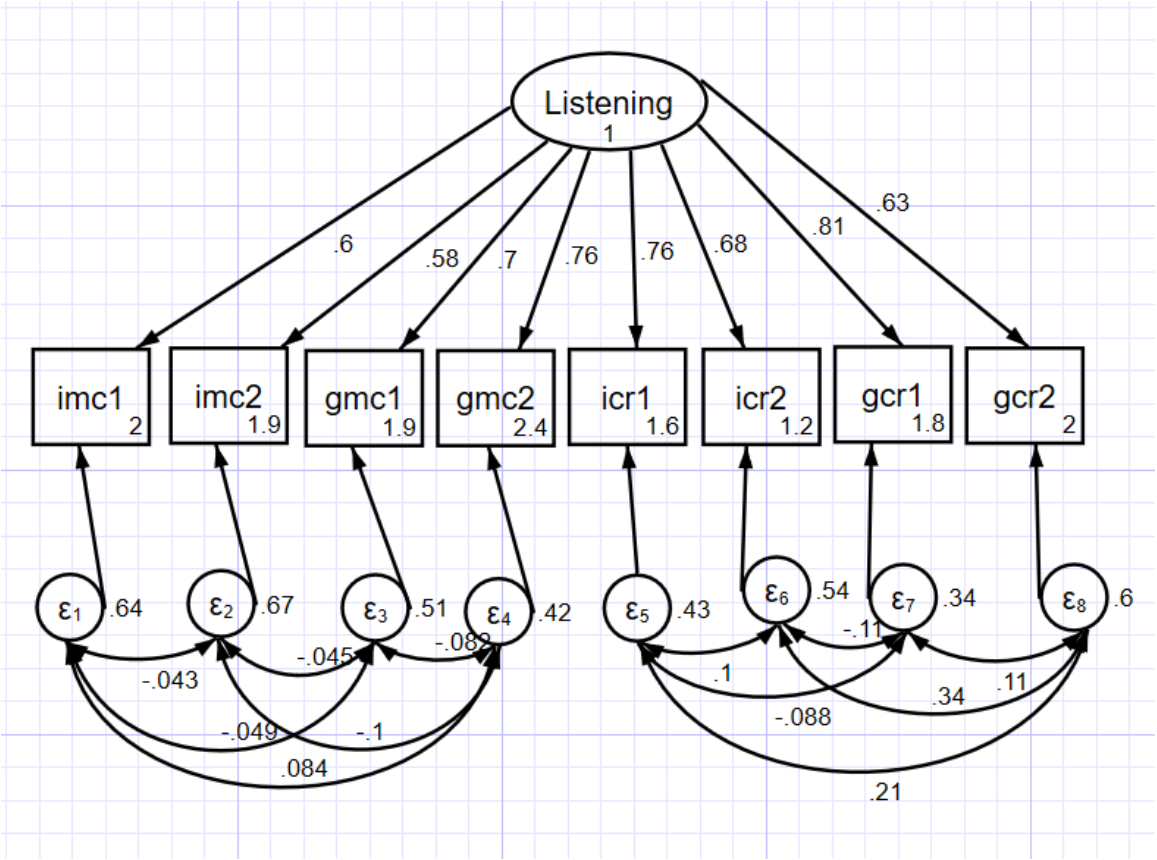


Figure 20. One Factor Model with Method Errors Correlated

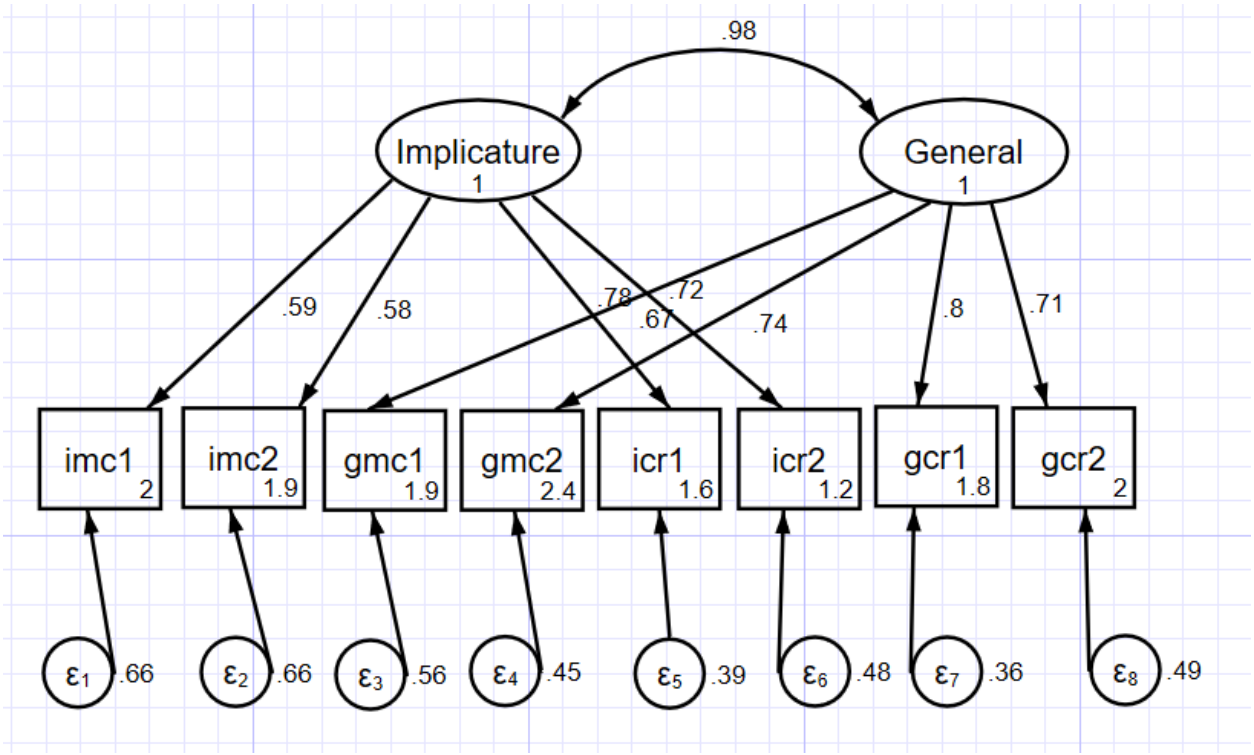


Figure 21. Two Factor Model with Method Errors Uncorrelated

The next two models that were run (Figures 21 and 22) examined the effect of separating the single latent variable of listening into two latent variables: one for general listening and one for implicature listening. As mentioned above, the TCI was designed to operationalize this conceptual separation, so this model will allow for testing of that designed operationalization. In the model in Figure 21, the method errors were not correlated, whereas in Figure 22, they were. The first value to look at is that the correlation between the two latent variables is extremely high: .98 when method errors were not correlated and .97 when they were. One would expect a high correlation because all items are testing listening, but when it is this close to 1.0 it is evidence that the latent implicature factor is indistinguishable from the latent general listening factor. The individual loadings from the indicators did not vary much from the single trait model either: In Figure 21 we see implicature loadings of .59, .58, .78, and .72 and general

loadings of .67, .74, .80, and .71. In the correlated uniqueness model (Figure 22) the implicature loadings were .52, .50, .89, .80 and general were .61, .66, .94, and .73. There is minimal difference from the one factor model loadings to these, and within the two different two-factor loadings, there is a slight improvement when errors are correlated for the CR indicator variables, but it is small.

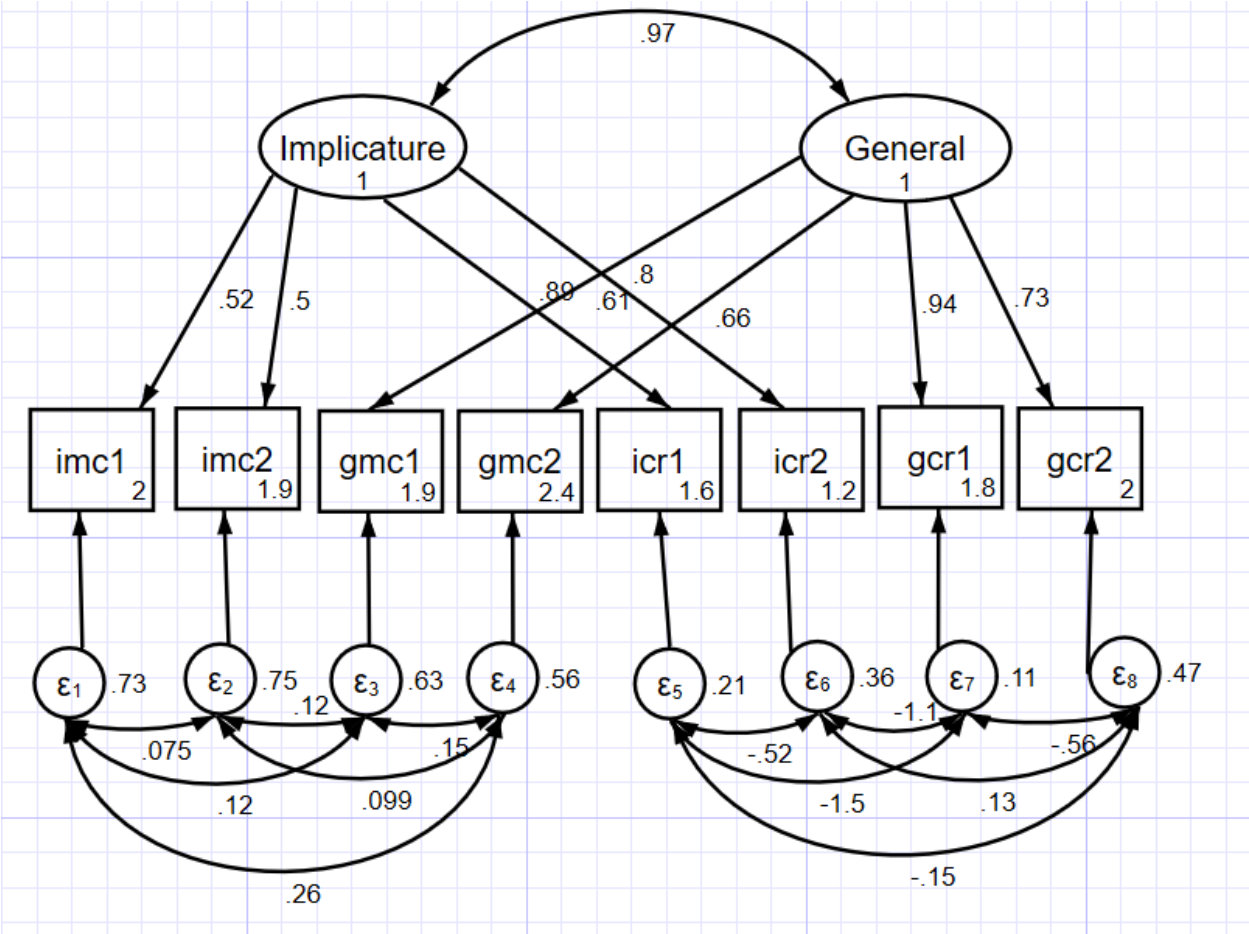


Figure 22. Correlated Uniqueness Model

Per Brown (2006), when employing the correlated uniqueness model, one can examine the "uniqueness" for each indicator as a proportion of the trait in question. The uniqueness is obtained by squaring the trait loading and subtracting that value from 1.0 and squaring it (Kenny, 2012; Kenny & Kashy, 1992 cited in Brown, 2006). Then, in a step that is reflective of Campbell and Fiske's (1959) original description of multi-trait

multi-method matrices, these uniqueness values are compared to the correlated uniqueness (method values). If all of the correlated uniqueness values are smaller than the uniqueness values, we see evidence that method is having a minimal effect at measuring the target trait. Table 17 below shows the uniqueness values from the model shown in Figure 26, and Table 18 shows the correlated uniqueness values

Table 17. Trait Loadings and Uniqueness Values for Correlated Uniqueness Model

	Trait Factor Loadings		Uniqueness	
Indicator	Implicature	General	Implicature Uniqueness	General Uniqueness
IMC1	0.52	NA	0.7296	NA
IMC2	0.5	NA	0.75	NA
GMC1	NA	0.61	NA	0.6279
GMC2	NA	0.66	NA	0.5644
ICR1	0.89	NA	0.2079	NA
ICR2	0.8	NA	0.36	NA
GCR1	NA	0.94	NA	0.1164
GCR2	NA	0.73	NA	0.4671

The picture from this model's uniqueness values shows fairly clearly that the multiple-choice method, included in the model via indicators IMC1, IMC2, GMC2, and GMC2, is not playing a stronger role than the target traits of general listening and implicature listening. The purpose of generating the uniqueness values for each trait (columns four and five in the table above) is to confirm that they are larger than all the correlated uniqueness values for the indicator variables, which are shown in Table 18 below.

Table 18. Correlated Uniqueness Values for Correlated Uniqueness Model

Indicator	**IMC1**	**IMC2**	**GMC1**	**GMC2**
IMC1	1			
IMC2	0.0754	1		
GMC1	0.2579	0.1159	1	
GMC2	0.0991	0.0991	0.1528	1
	ICR1	**ICR2**	**GCR1**	**GCR2**
ICR1	1			
ICR2	-0.521	1		
GCR1	-1.502	-1.134	1	
GCR2	-0.147	0.1312	-0.558	1

The uniqueness values for both implicature and general listening for multiple choice are all quite robust, 0.73 and 0.75 for implicature and 0.63 and 0.56 for general, whereas the correlated uniqueness values are all smaller, ranging from 0.075–0.26. This tells us that the method of multiple choice is not demonstrating much of an effect on those traits in this sample.

The picture is slightly less clear for the constructed response method (indicator variables ICR1, ICR2, GCR1, and GCR2). The uniqueness values in Table 17 are smaller than for multiple choice, ranging from 0.11 to 0.46 and one of the six correlated uniqueness values in Table 18 does surpass the smallest CR uniqueness value: the ICR2:GCR correlated uniqueness value is 0.13 and the general uniqueness value for GCR1 is only 0.116. However, it appears that even though the CR format introduces the additional skill of writing, the items are for the most part testing the target traits and are not inordinately testing ability to perform on the method. In fact, the presence of negative values for correlated uniqueness for the CR indicators is not a sign that the model did not converge (it did converge after 10 iterations), but rather a sign that the effect of method is quite small (Brown, 2006, p. 221).

The results of the four CFA models were further analyzed by comparing the fit statistics for each model (Table 19 below).

Table 19. Fit Statistics Summary for CFA Models

MODEL	**Chi square**	**RMSEA**	**CFI**	**TLI**	**AIC**	**BIC**
1 factor, errors uncorrelated	59.051	0.088	0.955	0.936	7457.803	7542.414
1 factor, method errors correlated	21.225	0.081	0.985	0.946	7443.977	7570.894
2 factor, errors uncorrelated	58.35	0.091	0.954	0.932	7459.104	7547.240
2 factor , method errors correlated	20.41	0.087	0.984	0.938	7445.163	7575.604

The statistics that were used to compare the models were chi-square, root mean square error of approximation (RMSEA), the comparative fit index (CFI), Tucker-Lewis index (TLI), and Akaike's information criterion, and the Bayesian information criterion (BIC). Following Hu & Bentler's (1999) suggestions for acceptable fit, for RMSEA, values of ≤ 0.06 were expected for a model with good fit and for CFI and TLI, values ≥ 0.95 were expected. The chi-square value was expected to be not significant, and AIC and BIC values were expected to be comparatively smaller for better fitting models. Using these criteria for adequate fit Table 19 shows a mixed picture. The chi-square values were smaller when the errors were correlated (Models 2 and 4), but all of them were significant. The RMSEA value did not drop below 0.06 for any of the models, but it was lowest for model 2, the one-factor model with errors correlated. CFI and TLI were adequate for model 2, although CFI and TLI were very similar in model 4. The AIC values were smallest, by a margin, for model 2, but BIC was smallest for model 1, although only marginally so. The overall picture by fit estimates appears to support what

has been discussed with the Rasch analysis section above: the listening test very likely was unidimensional, and while there was some effect of difference of method between multiple-choice and constructed response, it was minor.

5.5 Logistic regressions on CEFR level classification

To look at whether there is evidence that comprehending implicature is a higher ability skill, which was the focus of the third research question, with higher ability being defined in this case by CEFR level (as measured by an independent proficiency test, the MET), several logistic regressions analyses were run.

Roughly one-third of the participants (84 of 251, or 33%) provided their MET listening CEFR level or their MET listening scaled score, which was converted to a CEFR level based on the cut scores provided by Michigan Language Assessments (2012). Because the number of predictor variables in each logistic regression model ranged from only two to four, meaning the ratio of observations to predictor variables was between 40:1 and 20:1, the sample size was adequate to run the logistic regressions (Hinkle, Wiersma, & Jurs, 2003) used to answer the research questions pertaining to predicting CEFR level.

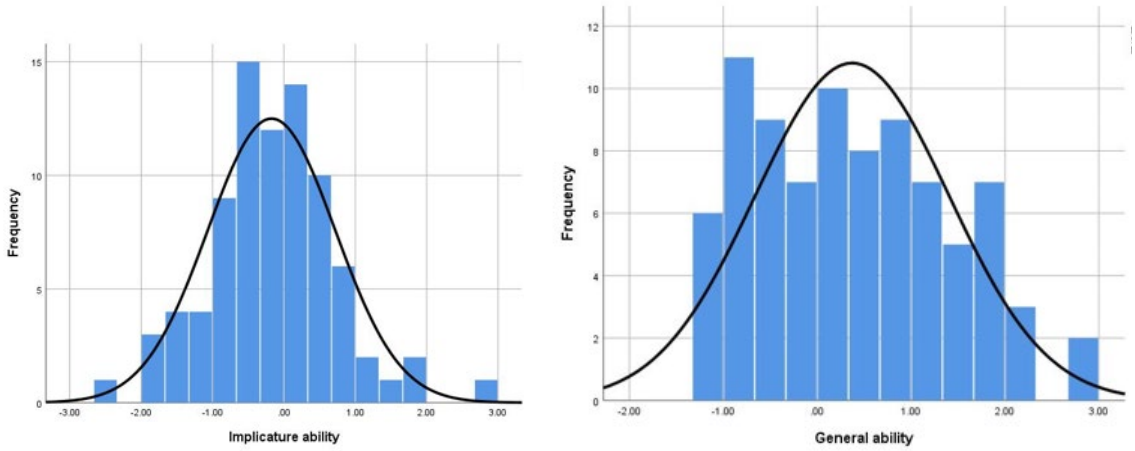
The shaded rows in Table 20 below show the descriptive statistics, skewness, and kurtosis for the subset of 84 participants that were part of the logistic regressions, which can be compared to the overall sample values in the unshaded rows. The working memory measures are included to show comparability of the 84-person subset on measures beyond performance on the listening test. The values of the 84-person subset are very similar to the overall group values, showing that they are representative of the overall sample.

Table 20. Descriptive Statistics for Participants with MET Scores

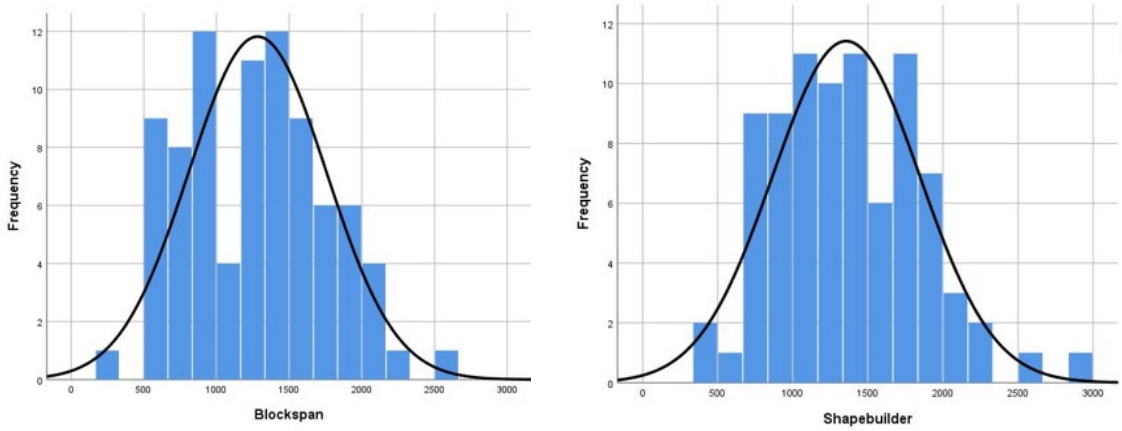
	N	Min	Max	Mean (std. error)	S.D.	Skewness (std. error)	Kurtosis (std. error)
Implicature ability*	251	-2.48	2.68	-0.17	0.86	.277 (.154)	.220 (.306)
	84	-2.48	2.68	-0.17	0.89	.181 (.263)	.835 (.520)
General ability*	251	-1.89	4.12	0.39	1.06	.417 (.154)	.025 (.306)
	84	-1.29	2.93	0.37	1.03	.353 (.263)	.686 (.520)
Shape-builder	248	66	2930	1362.32 (30.66)	482.927	.228 (.155)	.024 (.308)
	84	425	2930	1355.30 (52.7)	489.099	.512 (.263)	.249 (.520)
Blockspan	249	220	3000	1291.33 (31.1)	490.652	.684 (.154)	.649 (.307)
	84	300	2510	1282.98 (51.0)	472.432	.203 (.263)	.608 (.520)

*As reported in sections 5.1 and 5.2, implicature ability estimates were generated from performance on implicature CR and implicature MC items; general ability estimates were generated from performance on general CR and general MC items.

Figures 23–26 show the distributions for the subset of 84 on implicature ability (Figure 23), general abilty (Figure 24), Blockspan (Figure 25), and Shapebuilder (Figure 26):



Figures 23 and 24. Implicature Ability and General Abilty for 84-Person Subset



Figures 25 and 26. Blockspan and Shapebuilder Scores for 84-Person Subset

The subset of 84 were checked for outliers as well, and as the boxplots in Figures 27 and 28 show, no outliers (per the 3.0 IQR rule: the outlier indicated in Figure 27 and the two indicated for implicature ability in Figure 28 are per the 1.5 IQR rule) were seen in the working memory or listening ability score ranges.

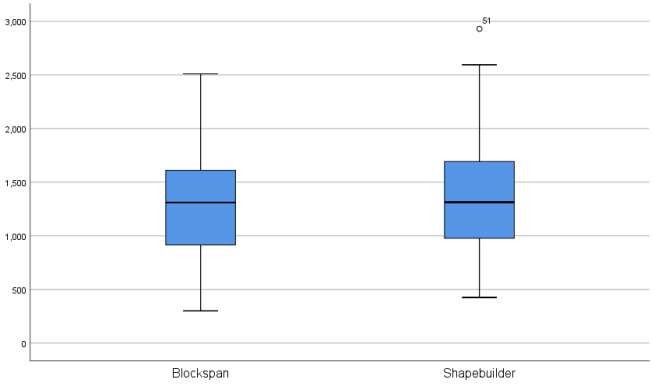


Figure 27. Blockspan and Shapebuilder Performance for 84-Person Subset

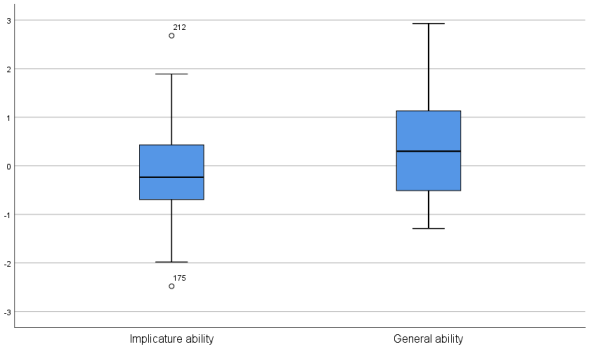


Figure 28. Implicature and General Ability for 84-Person Subset

After checking for normality of the distributions and for outliers, performance by CEFR group was examined. As the boxplots below show, the items on the TCI did separate C1 participants from below C1—both by implicature ability (Figure 29) and general abilty (Figure 30).

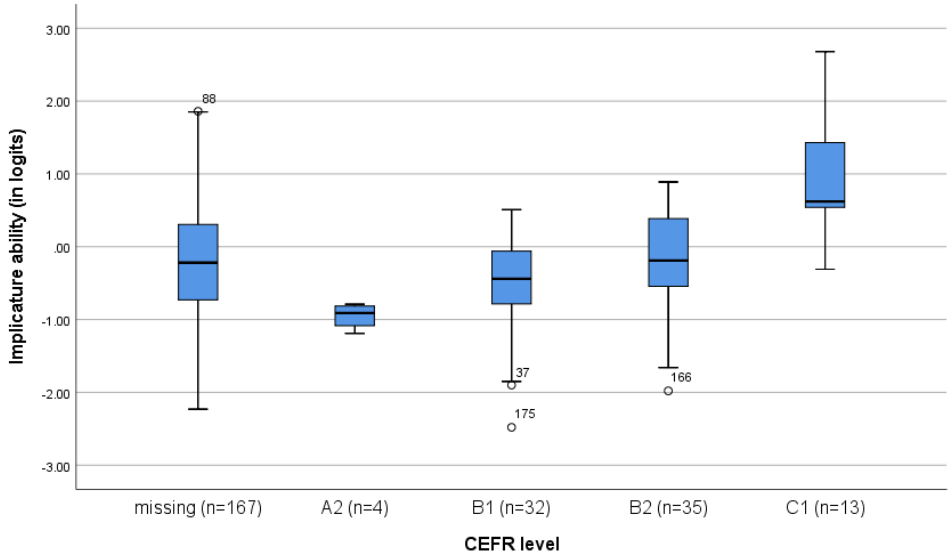


Figure 29. Implicature Ability of Participants by CEFR Level

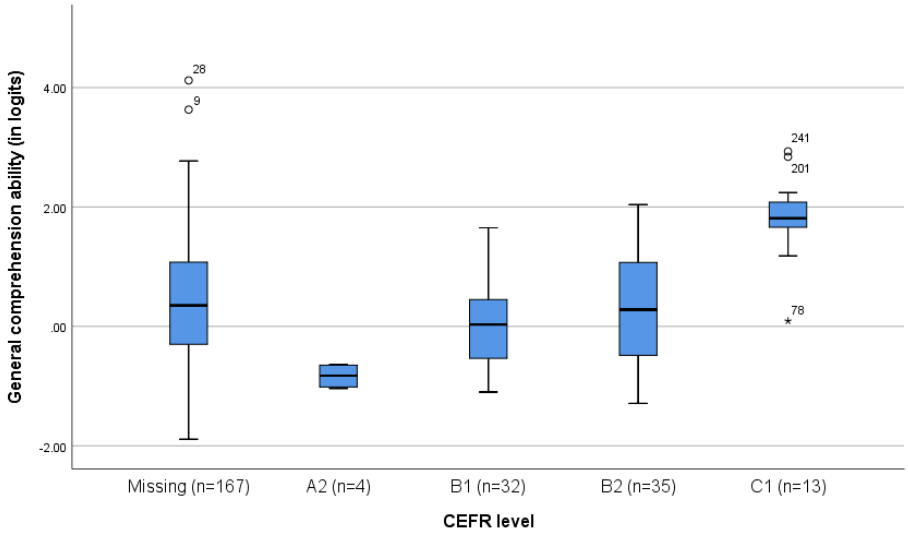


Figure 30. General Ability of Participants by CEFR Level

These boxplots also show that despite a desire to include participants in roughly equal numbers from B1, B2, and C1 levels, it appears that the large majority (n=167, the first box on the left) of the participants fell in the B1 and B2 range and that performance on the TCI did not seem to differentiate B1 from B2 level learners. This was supported by the logistic regression analyses described below.

5.5.1 Results of logistic regressions

Several logistic regression analyses were run with C1 classification as the dependent dichotomous variable (in the table below "1"=C1 and "0"=below C1, i.e., A2, B1, and B2 participants) and with general ability, implicature ability, Shapebuilder scores, and Blockspan scores as predictor variables. Analyses were also run with "1" defined as C1 and B2 and "0" defined as B1 and A2, but as the boxplots above indicate, B2 participants had considerable overlap with B1 participants in their performance on the TCI and there were no significant results via the logistic regressions. This makes theoretical sense too if it is assumed that implicit comprehension ability that is automatic is expected at C1, whereas it is developing or sporadic at B2 and B1. It is not the case that B1 or even A2 users of a language cannot comprehend implicature, but what is implied in the CEFR scales, and which seems to be supported by these results is that *regular* vs *sporadic* ability to comprehend the implicatures that appear in everyday conversation is an attribute of C-level language users—most likely because their automaticity with the language allows for the explicature processing step to happen in real time.

The first model that was run with the C1 vs below-C1 groups included all four predictor variables. The initial classification table (Table 21 below) shows that a model with no predictor variables included would be correct 84.5% of the time by simply predicting all test-takers were below C1—it would accurately predict 71 test-takers as below C1 but it would inaccurately place all 13 C1 test-takers as below C1.

Table 21. Classification Table for Participants without Predictors

Observed C1		**Predicted C1**	
	0	**1**	**Percent correct**
0	71	0	100.0
1	13	0	0.0
		Overall percentage	84.5

With the four predictor variables in the model, the Hosmer and Lemeshow test (Hosmer & Lemeshow, 2000) was significant at 0.011, with a Chi-square value of 19.758, indicating that the overall model did significantly improve the ability to predict C1 vs below C1. Of interest in a logistic regression analysis is by what percentage the model improves its ability to classify, and in this case it is a respectable 5% improvement, from 84.5 to 90.5 percent (Table 22).

Table 22. Classification Table for Participants with Predictors in Model

Observed C1		**Predicted C1**	
	0	**1**	**Percent correct**
0	67	4	94.4
1	4	9	69.2
		Overall percentage	90.5

Also of interest to the current study, which hypothesized a strong role for implicature ability, is the influence of the individual predictor variables. Table 23 summarizes those results from the first logistic regression analysis.

Table 23. Variables in Logistic Regression Model 1

							95% C.I. for Exp(B)	
Variable	**B**	**S.E.**	**Wald**	**Df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Implicature ability	1.789	1.128	2.515	1	.113	5.985	.656	54.651
General ability	2.175	.916	5.643	1	.018	8.801	1.463	52.952
Blockspan	.000	.001	.092	1	.762	1.000	.998	1.003
Shapebuilder	-.001	.001	.395	1	.530	.999	.997	1.001
Constant	-4.346	1.870	5.398	1	.020	.013		

In these results we see that general ability, rather than implicature ability, was the only statistically significant predictor at $p<.05$. Its Exponential B (Exp(B)) value, which is the exponentiated value of the B coefficient and is generally referred to as odds ratio in relation to logistic regression, tells us that for each increase in one logit of general ability, the likelihood of being classified as C1 (vs below C1) increases by a factor of more than

eight. The confidence interval values in the far right columns give us additional information that we can trust this result for general ability's role, as it does not cross 1.0. In logistic regression an odds ratio/Exp(B) value of less than one is an indicator of a negative relationship, an odds ratio/Exp(B) value of greater than one is an indicator of a positive relationship, and a value of 1.0 is an indicator of no relationship. If the confidence interval values do not include 1.0, it creates greater confidence in the output. In this case, the Exp(B) of general ability is 8.801 and the lower threshold is 1.453. In comparison we see the non-significant Exp(B) for implicature ability with a lower threshold value of 0.656 and an upper value of 54.651, which tells us that we cannot have confidence that that the positive odds ratio will be replicated with other similar samples.

The lack of significance for both WM measures was of interest as well, and although tests of multicollinearity did not show overlap between the two measures, their constructs are very similar, so two additional logistic regression analyses were run—one with general ability, implicature ability, and Blockspan as predictor variables, and one with general ability, implicature ability, and Shapebuilder as predictor variables.

The overall model for both analyses was significant (as determined by the Hosmer and Lemeshow test) and in both cases the classification accuracy improved over the 90.5% correct percentage that was seen in the first model (Table 22 above). For the model with general ability, implicature ability, and Blockspan as predictors, it was 91.7% classification accuracy, and for the model with general ability, implicature ability, and Shapebuilder as predictors, it was 92.9%. But as in the first model, neither WM measure was statistically significant as an individual predictor—the only predictor that was statistically significant (at $p<.05$) remained general ability.

The final model that was run included only general ability and implicature ability as predictors. This model also achieved 92.9% classification accuracy—the same as general, implicature, and Blockspan. Again, general implicature ability was the sole statistically significant individual predictor.

Table 24. Summary of Logistic Regression Models

Model (n=84 for all)	**Predictor variables**	**Significant predictor variables (p<.05) and 95% C.I. does not cross Odds Ratio of 1.0**	**Hosmer-Lemeshow test of model**	**Classification accuracy change**
1	Implicature ability, General ability, Blockspan, Shapebuilder	General ability (p= 0.018)	.011	84.5% -> 90.5%
2	Implicature ability[1], General ability, Blockspan	General ability (p=0.016)	.019	84.5% -> 92.9%
3	Implicature ability[1], General ability, Shapebuilder	General ability (p=0.016)	.013	84.5% -> 91.7%
4	Implicature ability[1], General ability	General ability (p=0.016)	.019	84.5% -> 92.9%

[1] Implicature ability significant at p<.10, but 95% CI crosses 1.0 for the Odds Ratio

What the fourth model tells us is that the working memory variables in the equation did not improve the classification accuracy beyond solely having general and implicature ability as predictors in the models (this was confirmed by running a logistic regression with only Shapebuilder and Blockspan, and the classification accuracy remained at 84.5%).

While these results are not what was predicted, they are informative, and upon reflection, not surprising. The prediction of influence of implicature ability was made on

the basis of existing language proficiency frameworks positing implicature and inferencing as higher ability, along with past research that has shown inferencing to contribute to difficulty (e.g., Kostin, 2004; Rupp et al. 2001). It appears that these analyses contribute to the assumptions inherent in the CEFR while also supporting past findings about inferencing contributing to difficulty.

As a way to employ a confirmatory step for the logistic regression findings discussed above, a linear regression analysis was also conducted on the MET listening scale scores of the 84-person subset of the sample who could provide those scores. The MET listening scores (scaled from 0–80) are what were used to place participants into CEFR levels, and served as a continuous (interval) dependent variable. Looking at this polytomous dependent variable was a way that the logistic regressions results for a binary dependent variable could be reinforced if not replicated. The same predictor variables were used as in the logistic regressions: general ability, implicature ability, Blockspan, and Shapebuilder. The overall model was significant, with an R^2 of .39 and as the regression output in Table 25 shows, the general ability predictor was significant while the implicature ability estimate was not.

Table 25. Coefficients in Multiple Regression on MET Listening Scores

	Unstandardized coefficients		**Standardized Coefficients**		
	B	**Std. Error**	**Beta**	**T**	**Sig**
(Constant)	52.848	2.599		20.333	.000
Imp ability	2.052	1.416	.205	1.448	.151
Gen ability	3.901	1.235	.451	3.159	.002
Blockspan	.001	.002	.030	.269	.789
Shapebuilder	-.001	.002	-.066	-.607	.546

To interpret the logistic and multiple regression output in conjunction, it appears that the source of difficulty in the TCI is the task itself, i.e., processing in real-time an exchange between two proficient speakers talking at a natural pace and making no linguistic accommodations, rather than the subskill that is the focus of the items. The boxplot of general ability by CEFR level (above in Figure 30) shows a distance from those who are at the C1-level from those that are below. This distance between C1 and those below is also seen in the implicature ability in Figure 29, but there is an even tighter grouping (save one outlier) of the C1 listeners according to general ability than by implicature ability. What can be hypothesized from that and the multiple regression is that general proficiency—ease and automaticity of processing lexical and syntactic structures is necessary to attain high levels of proficiency—and this ability is what allows ability for implicature to be more consistently applied. It also should be noted that the TCI general items probably more closely track with the listening section of the MET. The MET is a general proficiency test which probably includes listening for inference items, but they are not likely to be the majority of the items, and within the category of listening for inference, listening for conversational implicature would likely be an even smaller subset.

5.6 Multiple regression analyses for working memory

A large body of research has looked at the relationship between working memory capacity (WMC) and language processing, with much of it showing that WMC plays a role in L2 acquisition and performance. However, despite the many years of research looking at working memory's role in language acquisition, fewer studies have looked specifically at listening ability (e.g., Brunfaut & Révész, 2011; Vafaee, 2016; Wayland et al., 2013) and only one that this author is aware of that looked at listening and implicature specifically, Taguchi (2008).

In this case, the area of interest was the question of the continuing influence of WMC for higher-ability learners, so two working memory measures were administered to the participants. It was hypothesized that WMC would manifest in performance on implicature items, which are predicted to be more difficult than non-implicature items. Two short-term measures were also included in the study, but because of the problems with administration that were detailed above, they were not included in all analyses.

An additional rationale for including WM measures was to allow for exploring the hypothesis that one reason that WM continued to play a role for adult or higher proficiency L2 learners (which is generally not seen in adult L1 acquisition studies (e.g., O'Brien et al., 2006)) was that greater WM capacity supports the processing of aural language from which implicature had to be extracted from the explicature of an utterance (the three-stage processing that Sperber and Wilson proposed and discussed in detail in Chapter 2 above).

The research questions presented in Chapter 3 related to working memory are:

4a. To what degree does the role of working memory (as measured by both complex and simple tasks) influence performance when ability to comprehend implicature is measured by multiple-choice items?

Prediction: the null hypothesis will be rejected

4b. To what degree does the role of working memory (as measured by both complex and simple tasks) influence performance when ability to comprehend implicature is measured by constructed-response items?

Prediction: the null hypothesis will be disproven

4c. To what extent will working memory exert greater predictive influence on performance for multiple-choice items in comparison to constructed response items? Prediction: the difference will be non-significant (i.e., working memory plays a beneficial role in comprehending the implicature, but test method will not have an effect).

The results of the analyses show that the predictions made were not borne out, but in fact seem to present a coherent picture of the opposite scenario: working memory appearing to play less of a role with implicature items than with general listening items. These results will be discussed in detail below, after a discussion of the correlations.

The data were analyzed in three ways: 1) the data set limited to only test-takers with full memory measure item response rates, 2) the data set with only working memory measures, and 3) the working memory-only data set split by age. The influence of WMC was explored with five different independent variables: overall listening ability, implicature listening ability, general listening ability, MC listening ability, and CR listening ability. Prior to running the regressions, the distributions of the data and the possible multicollinearity were investigated, and correlation matrices and scatter plots were examined. Explanations and details on each data set are provided in turn, starting with the full data set.

The descriptive results for the memory measures based on the sample used in the regression analyses are provided in Table 26.

Table 26: Descriptive Statistics for Memory Measures

	N	items	Mean score	Median	Min	Max	SD	Reliability (Cronbach's alpha)
Blockspan (WM)	217	40	1291.33	1250	220	3000	490.65	.780[1]
Shapebuilder (WM)	221	26	1362.32	1312.5	66	2930	482.93	.836[2]
Forward digit span (short-term memory)	189	20	15.8	16	1	19	2.77	.682[3]
Backward digit span (short-term memory)	205	20	13.44	15	1	19	4.62	.857[4]

[1]reliability calculated from the 217 cases with 40 (all) item responses
[2]reliability calculated from 221 cases with 26 (all) item responses
[3]reliabilty calculated from the data set after participants with perfect scores were removed
[4]reliabilty calculated from the data set after participants with no correct answers were removed

Cronbach's alpha for both working memory measures was acceptable (values greater than .90 are considered good; values above .80 are acceptable; values above .70 are still usable, but below .70 is questionable). The reliability estimate for the backward digit span task was also acceptable (alpha = .857) although reliability for the forward digit span was not. It is the author's view that a methodological flaw was in play here. As described in Chapter 4, while these digit span measures were modeled on existing measures, their delivery was via LimeSurvey. This had the benefit of enabling a web-based delivery, which enhanced ability to enroll adequate participants for the Rasch and SEM analyses, but had the drawback of not restricting or monitoring responses in the way delivery through smaller scale experiment administration via computers with specific software programs downloaded on them (DMDX, E-Prime, etc.). As explained above, the result was skewed response patterns for the BDS and FDS. Because of its

skewed distribution, which could reasonably be attributed to an administration flaw, and its poor reliability, the FDS was not included in any analyses. The initial analyses did include the BDS; despite its skewed distribution, its reliability was acceptable. Furthermore, the correlation matrix in Table 27 below shows that the FDS scores did significantly and moderately correlate with both working memory measures, which one would expect due to STM being conceptualized as one component of working memory (with executive control being the other major component). The distributions for the working memory measures displayed none of the problems seen with the two short-term memory measures, and as expected, scores on the two working memory measures correlated positively at .59.

Table 27. Correlation Matrix for Memory Measures

	N	**Blockspan**	**Shapebuilder**	**BDS**
Blockspan	205	1		
Shapebuilder	204	.593**	1	
BDS	196	.454**	.445**	1

**correlation is significant at the 0.01 level (2-tailed)

An initial set of regression analyses were run to explore the predictive role of WM and STM (as measured by the BDS task) on a dataset of 195 participants (it included only the participants whose BDS scores fell between 1–19, which removed 10 people). The regressions were run with a direct method of entry for the three independent memory variables (i.e., all entered at the same time) on each of five different listening ability estimates that were detailed in section 5.1 above: overall ability, general ability, implicature ability, MC ability and CR ability. None of the analyses resulted in statistically significant models.

However, because the initial research questions were focused on working memory, and because administration of the STM tasks created some questionable results,

the five sets of multiple regression analyses were then run again without BDS scores in the model. That is, there were only two independent variables, Blockspan and Shapebuilder, and the results were quite different. These analyses were also run with direct method of entry (no hypothesis about differential influence of one WM measure over the other), but the sample size went up from 195 to 248, because there were 248 participants with full listening and working memory test results, versus 195 with full listening, working memory, and BDS results. Table 28 provides a summary of the five analyses with Blockspan and Shapebuilder as predictor variables.

Table 28: Relationship of Working Memory to Listening Ability Measures

Dependent variable	**Pearson Correlation with Blockspan[1]**	**Pearson Correlation with Shapebuilder**	**p-value (of model)**	**Blockspan Standardized Coefficients Beta**	**Shapebuilder Standardized Coefficients Beta**	**R^2**
Overall listening ability	0.127*	0.147**	.052	.060	.111	.024
General listening ability	0.131*	0.171**	.023*	.044	.144	.030
Implicature listening ability	0.108*	0.105*	.173	.069	.064	.014
MC listening ability	0.164**	0.163**	.016*	.104	.100	.033
CR listening ability	0.101	0.134**	.102	.032	.115	.018

[1]Pearson correlation for Blockspan to Shapebuilder with this sample was .603 (significant at <.01)

Obviously, the R square values for all the models and the beta coefficients for the WM measures across the models, whether statistically significant or not, show that there is little explanatory power of working memory on listening ability for this sample. However, the indications of some differential effect between implicature and general ability and MC and CR ability do provide some information. Namely, the lack of a statistically significant result for working memory's role with implicature—which echoes the lack of correlation found in Taguchi (2008)—but one for general listening, is counter to the hypothesis presented above that WMC is applied to processing of implicature. These results do seem comprehensible if one understands that implicature—although so prevalent and common in speech—draws on real-world knowledge in a way that possibly is less dependent on WM and more dependent on LTM.

The results for effect of WM on listening ability when divided by item-format are close to what was hypothesized—that there would be no effect (which would support the assumption that MC is just as credible a way of testing implicature as CR, when test-takers are not presented with a given interpretation), but there is still a difference. And if one is to give statistically significant results some credence (however small the R square) for the subskill analysis, it ought to be given that same credence here. But what we are possibly seeing is that in requiring test-takers to mentally construct their interpretation of the conversation and then write it (i.e., CR format), versus mentally construct their interpretation and match it to one of four given options (i.e., MC format)—the influence of WM on their comprehension is likely modulated. That is, in the CR format there is less reliance on WM but more from LTM as one must draw from syntactic, semantic, and orthographic knowledge in LTM to construct a coherent response. In sum, we see—as

minimal as it is—some evidence that is in line with past research about the role of WM in language processing. But when slicing listening ability by subskill we see some evidence that contrasts between two subskills and two methods, but in a minor way. But overall, the influence of WM on performance on this listening test, in light of the extremely small R-squared values, must also be described as minor, with the linguistic knowledge that participants brought to bear on the tasks, both in terms of their depth of vocabulary and knowledge of syntax, likely carrying the bulk of explanatory power.

Before moving away from the working memory analyses, a secondary, post-hoc look at the data was conducted that bears mention. While collecting this study's data, cleaning it, and preparing it for analyses, it was observed that age seemed a potential factor for the working memory measures. While a minimum requirement of 18 was set for this study, and roughly 87% of the participants were under age 30, no maximum age limit was set. In light of past research (e.g., Brehmer, Y., Li, S. C., Muller, V., von Oertzen, T., & Lindenberger, U., 2007) that has indicated that WMC capacity starts to diminish as people get older, a dichotomous age variable was created to separate participants aged 35 and older from those under age 35. The effect for WM did increase across all models, as is shown in Table 29 (note that BS=Blockspan and SB=Shapebuilder).

Table 29. Relationship of WM to Listening Ability Measures with Sample Subset

Dependent variable	**Pearson Cor-relation with BS**	**Pearson Cor-relation with SB**	**p-value**	**BS Standardized Coefficients Beta**	**SB Standardized Coefficients Beta**	**R square/ adjusted R square**
Overall listening ability	0.140*	0.203**	.007*	.032	.184	.042 / .033
General listening ability	0.147*	0.228**	.002*	.021	.216	.052 / .044
Implicature listening ability	0.118*	0.153*	.057	.043	.128	.025 / .016
MC listening ability	0.175**	0.204**	.004*	.084	.154	.046 / .038
CR listening ability	0.111	0.189**	.015	.001	.188	.036 / .027

However, although there was a greater effect across all the models, with nine of the ten ability-measure to WM measure correlations increasing and all five model's R square value increasing, the beta coefficients and R square values are still small, and these additional analyses only serve to reinforce the already known need for consideration of age in studies where working memory is a central focus.

Chapter 6: Discussion and conclusion

The ability to comprehend conversational implicature—operationalized as a speaker's intended meaning obtained beyond the surface features of the utterance—was investigated in this study by contrasting item format (multiple-choice and constructed-response) and subskill (implicature and non-implicature listening), and was analyzed with several different statistical methods, including Rasch, logistic regression, and CFA. The results are summarized in this chapter and then discussed in regard to implications for understanding the process of comprehending implicature and for using implicature items in language proficiency assessments, both in terms of a validity argument and in relation to the CEFR.

6.1 Research question 1

The first research question asked whether the correct answer choices, or keys, that are provided in multiple-choice implicature items are comparable to what the test-takers would generate themselves. That is, for the conversational implicature items that were created and administered, is there a clearly accessible inference that is available to a majority of listeners? This is an important question because of the continuing prevalence of multiple-choice testing in language proficiency assessments, and because there continues to be skepticism about the format's ability to test more complex types of understanding, i.e., beyond comprehension of main ideas or significant details. The results from this study showed that when multiple-choice items follow rigorous test development practices, the keys that are generated by test developers can be matched by what is generated by target-level test takers. The qualifier "target level" test takers is important here, however. The constructed-response results that were examined (discussed in section 5.3.1 above) to answer this question were from the top 10% of performers in

the sample, a decision that was taken because the independent proficiency measure used in this study showed that the bulk of participants were on the low intermediate/intermediate line (i.e., B1/B2 on the CEFR) rather than on the high intermediate/low advanced line (i.e., B2/C1), the latter being the level for which the items were designed. However, the ability of the top 10% to generate the item keys on 29 of the 30 implicature items is evidence that the tested keys created by test developers are reasonable representations of understanding that non-test developers would take from the exchanges.

Granted, the item's *stem* was generated by the test developers in both formats for this study. In future work in this area it would be intriguing to ask test-takers to simply generate a short "what was communicated in this exchange" type response (i.e., following Clark's (2007) "Model Comprehenders" step)--and then compare those responses to the selected stems and answer pairs generated by the test developers, which for the most part represented the test developers' understanding of what the salient or most important message of the short conversations was. In rigorous test development contexts, this is a step that is a formal part of the item development process: test developers listen to the selected or drafted aural input and identify the key message or most salient information before the items are created. This step enhances the ability to avoid testing trivial parts of the input and instead focusing on that which is most salient (and is particularly important for longer audio stimuli). The fact that keys were generated in the CR format by the participants in this study is noteworthy, particularly as there has been little work in this area for listening comprehension. As noted in the literature review above, more CR to MC comparison work (regardless of subskill) has been done with reading, and what has

been done with listening has been relatively small scale (Buck, 1991; Wu, 1998) and has tended to imply that expected keys will not be obtained when constructed by test-takers.

As a final comment on research question 1, the aggregate difficulties of MC format compared to CR format, which will be discussed further below, also lend support to the notion that MC keys are accessible. The absence of a stark mean difference in difficulty between the two formats allows the inference that learners who do well on MC items also do well on CR items—because it is more their listening that is being tested and less their test-wiseness, or skill with a particular test item format.

6.2 Research question 2

The second research question asked if comprehension of conversational implicature was a skill distinct from non-conversational implicature. This question was looked at in a number of ways with multiple analyses, including the Rasch results by subskill (section 5.3), an examination of principal components of item residual variance in Rasch (section 5.3), logistic regressions on individual implicature items (section 5.3.2), and a confirmatory factor analysis (section 5.4). All of these results converged to create a picture that when subskill type is the only factor that differs in the item specification (as was the case in the test used in this study), there is not a meaningful separation in the item types. Further, the fact that item specifications only differed by subskill was an attempt to avoid item difficulty being a confound in the study. That is, if a set of 30 implicature items are clearly more difficult than a set of 30 non-implicature items, it was hypothesized that the difficulty was due to subskill, and not vocabulary, dialogue length, or speakers, etc., which were controlled.

The Rasch difficulty results by subskill that were presented in section 5.3 showed that while the implicature items performed at an aggregate average of greater difficulty

than the non-implicature items, both in multiple-choice format and constructed-response format, the difference was not statistically significant. This lends credence to previous findings that implicature contributes to difficulty (Kostin, 2004; Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R., Mollaun, P., Nissan, S., Powers, D. E., & Schedl, M., 2008, etc.), but does not support a hypothesis that it is a different type of listening. The principal components analyses of Rasch residual variance, which provides a way of looking at groupings of subsets of items whose residual variance is correlating in a manner that is not expected, showed no evidence of implicature items clustering together. In fact, the results of the principal components analysis of residual variance on both MC and CR showed good evidence of the unidimensionality of the items.

The logistic regressions that were conducted on the probability of getting each individual implicature item correct were an additional set of analyses run to look at this question at the item level. The evidence pointed to the greater relevance of general listening skill, as the probability of responding correctly on 15 of the 30 implicature items was improved by test-taker's performance on general items, versus only six of the 30 implicature items showing an influence from test-taker's performance on their implicature items. These results would seem to indicate that general listening subsumes listening for implicature, rather than there being anything distinct about it. On one hand, this could be reassuring for test developers interested in testing implicature, where some have voiced concerns that implicature by its very nature of involving context and shared background knowledge makes it potentially problematic for including in standardized listening tests. The question of topic and shared background knowledge is one that must be addressed,

and will be discussed below in relation to the six items that were predicted by implicature raw scores.

The results of the confirmatory factor analyses, a widely used method for looking at the question of separability of constructs, also pointed to the unidimensionality of the listening test. While none of the fit indices for the four models reported on in section 5.4 were within the category of acceptable, the single factor model with method correlated by error had the best overall result. This implies that the test items were unidimensional and there was some shared variance by format. The model that was predicted to have the best fit, the correlated uniqueness model, which included two listening traits and errors correlated by method did not have the best fit, but an analysis of its correlated uniqueness values (Table 18 above in Section 5.4) gives evidence that not only was the method of multiple-choice format not something that was being tested, but also that constructed-response method did not play a role in performance.

As a final comment on the apparent lack of evidence for a separate implicature subskill, one of the key points to keep in mind is that listening comprehension involves a multitude of factors (Buck, 2001; Bloomfield, Wayland, Rhoades, Blodgett, Linck, & Ross, 2010; Rost, 2011) which interact and contribute to difficulty in different ways. But the findings here are still meaningful, because 1) the test development process employed here is very similar to what test developers are doing in the real world when they assign subskill tags to listening comprehension items on general proficiency tests, and 2) there continues to be demand from general proficiency test-users in the real world for subskill information. There may be ways to try to isolate listening for implicature even further, and some of those ideas will be discussed below. But in the meantime, what these results

seem to indicate is that listening for implicature subskill information seems to refer to a listening skill that develops at the same rate as non-implicature listening skills.

6.3 Research question 3

The third research question asked whether performance on the TCI would be differentiated by CEFR level. Rather than addressing the construct separability topic, this question focused on whether there was an increased difficulty (or success) with implicature items in relation to a commonly used framework of reference. The CEFR is of interest here, primarily because it is a widely used and helpful tool for comparisons of language proficiency across languages and contexts, but also because its listening scales strongly imply that mastery of comprehension of implied meanings is a "C level" skill. While there is no assertion made by the CEFR (or the author) that lower-level language learners cannot comprehend implicature, it is asserted by the CEFR, and supported by these results, that being able to comprehend implicature with regularity with everyday dialogues is something lower-level learners cannot do—when the language that the implicature is embedded in is not simple or modified. By design, the conversations and items used in this study were created for high intermediate to proficient listeners. That is, they were written, reviewed, revised, and recorded to be naturalistic conversations employing natural syntactic structures and discourse organization all delivered at a normal speaking speed by proficient speakers of English. The question of whether CEFR level (as proxy for general listening proficiency) played a role in performance on the test was answered quite clearly by the logistic regressions reported in section 5.5 (and supported by a multiple regression that was run with participants' MET listening scale scores as the dependent variable).

At this point, it is useful to provide some background on the MET listening section. Per Michigan Language Assessment's website, the MET listening section aims to measure a learner's "ability to understand conversations and talks in social, educational, and workplace contexts" (Michigan Language Assessment, 2019). This is done via three types of four-option multiple-choice tasks: listening to short dialogues, listening to extended conversations, and listening to short talks. The latter two tasks are multi-item tasks, meaning that test-takers will respond to three or four items after listening to the extended dialogue or the talk. The first listening task on the MET is very similar to the task used in this study: short dialogues following by a single item. The make-up of the MET listening section is of interest because it is unlikely that a preponderance of the items, or even half of the items, on it are meant to assess implicature. However, even without access to the test's specifications, the MET sample test form made available on the website gives an indication that at least 25% and probably more, of the items assess implicature to varying degrees: prediction items, suggestion items, and rhetorical function items are all evident from the stems in the sample test form. The point being that implicature items, of varying degrees, are including on the MET listening section in a non-trivial way, and therefore are contributing to the overall score and CEFR placement of test-takers. Thus, the predictive analyses from participants' TCI results in this study to their MET listening level are not necessarily biased towards general listening versus implicature listening. Rather it seems to be an open question and of interest whether the implicature items or the general items from the TCI are more predictive of higher-level MET performance, and whether that influence is seen from the B2 level or from the C1 level.

Returning to the research question, when the logistic regression was run with a dependent variable of C1 vs below C1, the model was significant and classification accuracy was improved when performance on the TCI (implicature ability and general ability) were included as predictors in the model. When the C1 and B2 level participants were grouped together, those results did not appear. What this tells us is the type of items that were designed for the TCI, both general listening and implicature listening, were effective in separating the advanced learners from the intermediate ones. The answer to the research question predicted that the implicature items would be better at this separation, but in fact while both types of items were effective predictors, the general items were more effective. Therefore, unless implicature is an important part of the listening construct for a given test use, performance on challenging general listening items will likely give test developers adequate information about whether or not test-takers will also do well on implicature items. The question of implicature items and test use will be discussed further below.

6.4 Research question 4

The final research question focused on the potential role of working memory with listening test performance by subskill and format. Multiple regression analyses were run to look at this question, and when there was only working memory in the model, and not the short-term memory scores that were collected, a small role was detected for working memory in regard to performance on the general listening items but not on the implicature items. This finding is in line with past research that generally sees working memory play a predictive role in L2 language performance, but the role for general listening ability and not implicature is the opposite of what this study predicted. The lack of a predictive role for implicature ability seems to indicate that the explicature step that

is necessary for reaching an implicature (per Sperber and Wilson), and the ability to connect context and topic to long-term memory, is happening outside of WM. It would seem that the role of WM for general listening found in this study—although quite weak—is in line with past research about it playing a role in the automaticity of processing that is necessary for all listening comprehension.

However, the weakness of the relationship between WM and general listening merits further exploration. One avenue was a consideration of the age of the participants. This study included no participants below age 18, but there was no cap on age and there were participants in the sample in their 40s and 50s. However, as reported in section 5.6 above, while the analyses with WM on a subset of the sample (only those aged 18–35) did show a stronger predictive relationship for WM, it was only marginally stronger. Another possibility for the weaker-than-expected relationship was the target level of the tasks. As reported above, the 60 items were created to require automatic processing of natural language when no accommodation for grammatical structure and minimal accommodation for lexical frequency were made. Thus, they were designed to be at the B2 to C1 level of the CEFR. It is possible that with more B1 participants in the data than hoped for (reported in section 5.5), there was some restriction on the range of listening ability that those participants could show, which would lessen the ability of WM to show its influence. Additionally, another step that could be taken with this dataset to explore the role of WM further would be to create a composite score from the Blockspan and Shapebuilder scores. Their relationship to each other was moderately strong and significant (.59 Pearson correlation), and the creation of a composite variable to include

in regression or SEM analyses may more clearly show the potential influence of working memory on listening outcomes.

But the findings with working memory contributing (slightly) to performance on general ability items but not with implicature ability items also lends support to the notion that what is happening with comprehension of conversational implicature is happening beyond language and language ability. That is, the literature shows consistent findings for WM and language processing and here there is correlation for items that were designed to try to test language alone and no correlation for items that were designed to test language *and* the ability to comprehend implicatures. It leaves open the possibility that if there is a distinct ability to comprehend implicature, it is distinct from language, that it may be the case that items tapping implicature "require different thought processes than the basic understanding questions" (Enright, et al. 2008), but that those "different thought processes" are entirely unrelated to language. There is potential that they are tied to non-linguistic reasoning skills, or perhaps to a type of empathy or "ability to predict" skill that anecdotally is attributed to successful politicians or salespeople. It could even be tied to the non-linguistic concept of "savviness" that Kramsch (2009) attributes to language learners who have higher levels of intercultural awareness than other language learners and therefore greater ability to communicate successfully. These possible variables could be investigated further—if there are reliable measures of them—in conjunction with a reliable measure of implicature such as was employed in this study.

The small predictive role of WM that was seen for MC ability but not appearing for CR ability provides a small degree of evidence (also seen in the CFA, where models with correlated errors by method had better fit) that while aggregate performance by item

format performance is comparable, there is a slight differential effect for format, which makes sense considering how different these formats are, involving an entirely different performance: thinking of an original response and typing it, versus matching one's response with the options given.

6.5 Implications for understanding conversational implicature

One component of Buck's default listening construct is understanding inferences that are "unambiguously implicated" (p. 114, 2001). But whether or not the information is ambiguous, as is made clear by research in the field of pragmatics, depends to a great extent on context and background knowledge. This obviously poses a challenge for developers of standardized tests who wish to test implicature. This challenge—the need to avoid ambiguity—is given high priority during item development, with part of the review process for all reviewers involving a step where they ask themselves if an item's topic is likely accessible to a large majority of the target learners. It can obviously never be unanimous, but the hope is that by including that step during multiple reviews, and by considering statistical performance from field testing, the role of topic—for a general proficiency test—will be minimized. But it can be impossible to rule out, and as was mentioned above in section 5.3.2 where the results of the logistic regressions on all implicature items were discussed, it may be the factor that contributed to six of the 30 items being predictable by a listener's ability to process implicature and not their general listening ability alone.

As mentioned above, the review process for the listening test used in this study included judgments on the accessibility of topics as a normal and expected step. For the most part, that normal and expected step appears to have worked: the statistical results of item analyses showed that for only two of the thirty items, when taken in conjunction

with a review of the content, the topics were not accessible. These were items LIM_03 (an ambiguous re-casting of a question in an unknown context, which was flagged by one of the three independent reviewers but was included in the final measure) and LIM_56 (topic of borrowing a snow shovel requiring interpretation by test-takers who likely never shoveled snow). Neither of these items, however, were part of the group of six items that were predicted by implicature scores in the regressions. Those items were LIM_15, LIM_16, LIM_21, LIM_23, LIM_24, and LIM_60 (shaded in Table 30 below).

Table 30. Post-hoc Summary of Implicature Item Topic Accessibility

Item code	**Topic**	**Accessible?**	**Item code**	**Topic**	**Accessible?**
LIM_01	Recap of a trip	Yes	LIM_16	Slow IT dept	Borderline
LIM_02	Taking a break at work	Yes	LIM_17	Number of quizzes	Yes
LIM_03	**Student behavior**	**No[1]**	LIM_18	Evaluating boss	Yes
LIM_04	Evaluating a book	Yes	LIM_19	Cost of football tickets	Yes
LIM_05	Number of staff	Borderline	LIM_20	Progress on research project	Yes
LIM_06	Volunteering	Borderline	LIM_21	Grocery shopping	Yes
LIM_07	Study group	Yes	LIM_22	Evaluating apartment	Yes
LIM_08	Kayaking	Borderline	LIM_23	Visa for conference	Borderline
LIM_09	Stolen bike	Yes	LIM_24	Painting garage	Borderline
LIM_10	Social event at work	Borderline	LIM_55	Cost of book	Yes
LIM_11	Evaluating dentists	Yes	**LIM_56**	**Snow shovel**	**No[1]**
LIM_12	Attending rally	Borderline	LIM_57	Evaluating exam	Yes
LIM_13	Evaluating teacher	Yes	LIM_58	Noisy apartment	Yes

LIM_14	Cost of city hall	Yes	LIM_59	Parking car	Yes
LIM_15	Evaluating consultants	Borderline	LIM_60	Recycling sign	Borderline

[1]Post-hoc determination of inaccessible topic (made on basis of reviewing statistics in conjunction with content)

Table 30 shows a secondary, or post-hoc, consideration of how accessible each topic was. The term post-hoc is used because all were deemed "accessible" before being put on the test. "Accessible," in the context of this study meaning it would be reasonable to expect that someone 18 or older would have either experienced directly or have heard of/observed others experience the topic of conversation to such a degree that it was familiar enough for them to understand—*or* that sufficient context was provided in the input to allow them to understand even if it were unfamiliar. However, in the wake of the author proctoring this study's testing sessions and having a better sense of the population, ten of the items were deemed to be potentially "borderline" in terms of topic accessibility in this post hoc review step. Note that the items were developed with an intent to mirror a general standardized language proficiency test, and could have been administered domestically in the U.S. or in any country with English language learners; i.e., there was no knowledge or expectation of where test-takers would live or what their first languages or cultural knowledge was while the items were being developed—as is often the case with standardized general English language proficiency tests.

The majority of the items that were classified as potentially borderline accessible were those that centered on workplace office interactions, which the large numbers of test-takers in the study sample who were in their early twenties or younger may likely have had limited exposure to. But there were also items around kayaking (LIM_08) and recycling signs (LIM_60) that seemed borderline during the secondary accessibility

consideration. However, it should be emphasized that these ten items were not marked (as a group) in any item performance statistical category, in terms of discrimination, difficulty, or fit, beyond the fact that five, or half of them, were predictive by raw scores on implicature items (in multiple-choice format). It should also be noted that five of the six that were predictive by raw scores on implicature items were in this borderline accessibility category. The hypothesis being that here there seems to be a thin line of evidence that items where the topic and context might be slightly ambiguous or potentially unfamiliar can be more answerable by some test-takers who have greater ease with making implicatures (as determined by their raw score sum on other implicature items). It is a hypothesis that can only be posited cautiously, but is one that merits closer investigation via a retrospective think-aloud: can the majority of test-takers identify speakers and their relationships to each other, and topic if not given in these types of short dialogues? How does that ability differ from simply making the inference-- extracting the message-- that is expected by the test developers? Comparison of initial performance in test-like conditions with qualitative responses afterwards could provide answers in those areas.

Another way to consider these six items is to investigate if in some way they contained more of a focus on the pragmatic stance of the speakers rather than the intended meaning of their utterances. This is a distinction made by some researchers (e.g., Purpura, 2004), and although it is difficult to operationalize in standardized testing contexts, it is worth considering. Essentially, the idea is that below the intended meaning—the implicature—there is also a pragmatic meaning—which may not map directly with the intended meaning. As has been discussed above, because these

conversations were followed by a stem created by the test developer, an intended meaning is what was being sought from the participants—and this is very clear from stems such as "Why is the woman worried?" (in LIM_15) which points the listener to the fact of the woman already being worried and asks them to process the source of the worry. LIM_23's stem similarly points to processing the intended meaning of the language ("What is the woman's concern?") as does LIM_24: "Why does the woman mention tomorrow's weather?" However, when stems are phrased as "What does the woman imply about the cherries?" (LIM_21) or "What does the woman imply about the tech department?" (LIM_16), there is the potential that the speaker's stance, beyond the language, is what is being implied. I.e., for LIM_16, that the woman is not only implying that the technology department "doesn't work quickly," when she says one will get a quick response from them but "…getting something resolved is a different story," but that she is also conveying that they are terrible, or incompetent, or the worst workers in the company. However, such a secondary "beyond intended meaning of the implicature" responses did not appear in the open-ended CR responses, so the link between prosody of the delivery and direction of the stem did seem to point participants to the intended meaning of the implicature—which was the intent of the study, taking the view that "implicatures are textbook cases of pragmalinguistic items, and pragmalinguistics is the language-facing part of pragmatics" (Leech, 1983 quoted in Roever & Taguchi, 2017). The intended meaning of implicature is what is currently assessed in many listening proficiency comprehension tests and was the focus here. The additional pragmatic layer is certainly of interest, and could be probed with listening items as designed in this study if they were administered with both a test developer's stem and a more open "what

message was conveyed" prompt (a la "model comprehender" approach described above). In other words, this work could be extended by administering the implicature items in four conditions: 1) with the original prosody and a targeted stem, 2) with the original prosody and an open stem, 3) an alternate prosody with same targeted stem as condition 1, and 4) an alternate prosody with the same open stem as condition 2. This would allow for a more fine-grained exploration of how test developers' choices of stem override, or interact with, the prosody of a dialogue and interpretation of intended meaning and salience of pragmatic stance.

But to link this study's results back to Sperber and Wilson's relevance theory, the framing here was investigating whether or not the listeners could determine the *relevance* of the utterance—i.e., the conversational implicature. Whether the items focused on intended meaning versus an additional pragmatic meaning was not the focus; the ability of the listeners to obtain the relevance of the utterance was.

6.5.1 Role of item format

Some of the differences between MC and CR performance (as shown in Table 10 in section 5.3.1) merit further discussion as well, particularly in terms of implicature degree. As stated in the introduction, one of the practical considerations that drove this examination of comprehending implicature was to look at whether the constructed response format had any benefit over the more commonly used multiple-choice format. In aggregate, the multiple-choice versions perform well enough to be a defensible format. When comparing MC to CR formats, we see that the MC format is not made considerably easier by "giving test-takers the options" versus requiring them to generate their response as stronger "proof" of them having come to that understanding through their comprehension processes themselves. But at the item level for implicature items, there

were some large differences between difficulty in MC and CR format for some items, with the trend being that the CR format was more difficult. Some of these items are analyzed below, with the goal of gleaning some insight about how test-takers may have interpreted the implicature items they were presented with.

The implicature type of LI_04 (full MC content in sample 6.1 below) was classified as indirect criticism and the degree was classified as "low." It was judged to be low because although the general context/domain (a work situation) is not provided, the specific context/topic (staff handbook) was explicitly provided in the first turn and no competing topics are introduced. There are also two clear cues to support the implicature being tested (that the handbook is old): "looks like it was printed in 1990" and "it does need updating."

(6.1) M: Sorry to bother you, Karen, but I have a quick question: Is this the current staff handbook?
W: The one with the yellow cover? Yeah, that's it.
M: This looks like it was printed in 1990…
W: Maybe not quite that far back, but it does need updating.

What does the man say about the handbook?

It is old.*
It is confusing.
It looks ugly.
It was difficult to find.

The fact that performance of this implicature item in MC format was on the easy side (Rasch difficulty -1.35) was not surprising. What was surprising was the greater than 1.4 logit difference in difficulty for the CR version. One strong possibility for this issue is mechanical: the use of the word "say" in an MC stem for an implicature item is not problematic, but for a CR version, it is. It pushes test-takers towards the literal response "it was printed in 1990" versus the implicature response "it looks old." MC items would

not include among the distractors a literal repetition of what was "said" if the goal is to test implicature. Many of the test-takers responding to the CR version of this item responded with a literal "it was printed in 1990," which resulted in only partial credit. This makes clear that item stems for a CR format need to be crafted specifically for the format. For example, "What is the man's opinion about the handbook?" would be an improved stem for an item based on the LI_04 exchange that is assessing implicature regardless of the item format.

LI_18 was another item with a large difference between CR difficulty and MC difficulty:

(6.2) W2: Do you think George is a good boss?
M1: Umm. [*hesitantly*] That's a tough question.
W2: I mean, what do you think about the way he talks to the employees?
M1: I really don't have a strong opinion about that.

Why didn't the man answer the woman's questions?

He doesn't want to talk about the boss.*
He doesn't know who her boss is.
He doesn't understand her.
He doesn't have time.

The implicature type in LI_18 was classified as "relevance/avoidance" and its degree as "high." This is an example where the high degree of implicature judgment (based on the general context not being provided and the fact that more than one implicature is available to the listener) was borne out in the CR responses, whereas, the MC results showed the item to be quite easy (-1.47 difficulty estimate). The CR responses showed a range of comprehension and participant ability to interpret the stem in relation to the exchange. This item appears to be a case where the response to a pragmatic implicature being supplied by the test developers in an MC format is probably more fair; many test-

takers got only partial credit for simply rephrasing the final utterance as "He doesn't have an opinion." Here the MC distractors can be reasonably eliminated, which is evident from its acceptable Rasch infit (0.84) and strong classical discrimination (0.56 and 0.67 in the two test forms where it appeared as an MC item).

Item LI_21 (shown in sample 6.3), which was one of the six items predicted by implicature raw scores discussed above, was also pre-judged as having a low degree of implicature. The specific context (grocery store/food market) was provided via multiple cues (asking about price of cherries, etc.), so it was predicted to be on the easier side.

(6.3) W: Excuse me, how much are the cherries?
M: Just seven-ninety-nine a pound, ma'am. And just picked yesterday.
W: Seven-ninety-nine... I remember when they used to be seventy-nine cents a pound.
M: Sorry to say, ma'am, but those days are long gone.

What does the woman imply about the cherries?

They are expensive.*
They are several days old.
They will not sell well.
They don't look delicious.

This was borne out in MC format (-1.74 difficulty), but in CR format it was medium difficulty (-0.19). Many participants got partial credit for saying "They were cheaper in the past," but that is not what she implies. She knows they were cheaper in the past. She is implying that the cherries are expensive *today*. In MC format more than 80 percent of participants were able to identify that implication, but when forced to produce it themselves only roughly 50 percent did so. What this seems to indicate is that MC items can be more effective in obtaining an indication of listening comprehension—the CR response gets clouded by the question of how much test-takers think they need to say. Those that opt for brevity and write "they were cheaper in the past"—which is very close

to the answer (hence receiving partial credit), would very likely be able to provide, upon additional probing, the implication being sought. Which would seem to lend support to the traditional view voiced by test developers who employ MC items that the format lends itself to efficiency and clarity.

While the majority of discrepant difficulty items fell in the category of CR difficulty being greater than MC difficulty, there were some instances where MC difficulty was greater than CR difficulty. Looking at these items in detail can also provide insights for informing item construction.

(6.4) M: Are you coming to the rally on Sunday? It's to protest the plan to knock down Ferber Library.
W: I don't know. It's a nice idea, but ultimately, they're gonna do what they're gonna do.

What does the woman say about the rally?

She doesn't think it will be effective.*
She won't have time to attend it.
She doesn't understand its purpose.
She disagrees with the people organizing it.

The implicature type for item LI_12 (sample 6.4) was classified as "indirect criticism" and its degree as "medium." The general context is not provided, and although the specific topic (attending a rally) is provided, there is only one cue in the exchange to support the tested implicature (that the woman does not think the protest rally will be useful/effective). In CR format, the item did seem to be "medium" difficulty—the difficulty estimate was 0.18 logits. However, in the MC format, it proved to be quite difficult, with a Rasch difficulty estimate of 1.44. Its fit values, point-measure correlation, and classical discrimination (>0.50 for 41 test takers were all strong, pointing to an item that "worked" in the MC format, but here, the effect of distractors is apparent.

They drew medium and lower level test-takers away from the key. In the CR format where distractors were not presented, facility was higher. (And as with LI_04 above, the CR version of this item would have likely proved even more accessible if rather than "say," the stem was written as "What is the woman's opinion about the rally?".) So here we see a case where an item's difficulty is increased by presenting distractors that are very attractive, rather than the difficulty of the input or question. But whether this would be revised after field testing for use as an MC item is highly unlikely.

In terms of improving understanding of implicature, what this qualitative analysis of several items, coupled with the aggregate results, seems to show is that degree of implicature as judged by the test developer has minimal predictive value. Consideration of a degree of implicature value that is linked to whether general or specific topics are given, and how many support cues are given, can likely improve test developers' ability to avoid testing implicatures where the degree is too high to be accessible. However, we see that matching degree with performance (i.e., an assumption of low degree leading to easier items and high degree leading to more difficult items) can vary greatly for particular items depending on the format. And when considering format at this granular of a level, MC format seems not only defensible but also more fair—assuming the wording of the stems is carefully considered. One can create MC implicature items where the "implicatures" are accessible to the participants (i.e., not forcing them into an implicature they don't see) and also avoid scenarios with CR items where it seems like some test-takers are getting partial credit because while they may have understood the full pragmatic intent of the conversation, they only produced the literal message in their responses.

To link back to the potential concern mentioned in the Research Question 3 section above that there is too much involved in listening comprehension to be able to say from these data that implicature is not a meaningful factor, even though the design of this study has narrowed the factors from past studies, further narrowing may be possible, and potentially may lead to uncovering an implicature subskill. In the case of this study, a considerable number of variables were more controlled than in past listening subskill research. Here, all participants were the same L1, all the items were created by the same item writer and went through the same review process. Furthermore, the comparison of implicature to non-implicature was done with a balanced and sizable number of discrete, or single, items, rather than via multi-set items, where three or more items are assessing lengthy aural input and one of the three or four items is an implicature and the rest are not. But granted that the narrowing in this study was done to reduce the potential effect of additional variables, further refinement could be made to try to truly isolate comprehension of implicature. For example, a counter-balanced test composed of thirty short two-turn conversations that are turned into two separate items where one response is explicit and one contains implicature (see samples 6.5 and 6.6) could help explore this question in an even more granular way.

(6.5) W: How did you like the lecture, Kevin?
M: It wasn't very interesting. The speaker didn't say anything new.

Question: What is the man's opinion of the lecture?

(6.6) W: How did you like the lecture, Kevin?
M: I can think of a dozen things that would've been a better use of my time.

Question: What is the man's opinion of the lecture?

Based on the results of the current study, sample 6.6 would likely be more difficult than 6.5, but when language proficiency is controlled, there will likely be no meaningful difference. Both conditions would likely be predicted by language proficiency and both would likely load to a single listening variable. (Although for an investigation that is constrained to dialogues of only two turns each, the syntactic features of the rejoinders would need to be examined closely; in the examples in 6.6 and 6.7, a far more syntactically complex response is evident in the implicit response compared to the explicit response.) A format like this would also be a step away from the interactional authenticity that slightly longer dialogues provide, or the type of listening tasks that appear on proficiency tests. However, probing the distinction in this way and seeing those predictions borne out or not would even more definitively point to implicature ability as being indistinguishable in terms of language from general language ability. It may be worth the further exploration, particularly if a third condition could also be composed to include common semi-verbal or nonverbal responses that map to the explicit and implicature responses; see the exchange in 6.7:

(6.7) W: How did you like the lecture, Kevin?
M: Uggh. [or semi-verbal "Meh."]

Question: What is the man's opinion of the lecture?

Such an investigation would be of interest is seeing how different proficiency levels are able to comprehend condition 3 and to the extent that conditions 1 and 2 responses correspond to the findings of this study. But returning to this study, the findings appear to point to implicature comprehension being another aspect of comprehension that proficient language users are able to do in real time. Implications of this understanding of

implicature comprehension for language proficiency test development and use are discussed below.

But before turning to test development implications, which again, was the context within which this study was centered, the discussion above leads to a natural question about potential implications for L2 pragmatics research in general. On one hand, the findings provide additional evidence to support the view that "processing implicature requires advanced proficiency" (Taguchi & Roever, 2017, p. 139). While the findings seem to support, from an *assessment* perspective, that assumptions can be made about implicature comprehension ability based on non-implicature listening ability, there are no assertions being made about lesser importance of drawing learners attention to pragmalinguistic forms and sociolinguistic principles of the language they are learning. However, more specific implications for L2 pragmatics may be drawn from further investigation of the constructed-response data gathered in this study. One possible avenue of research is to focus on learner groups (rather than items) in order to explore 1) the degree to which learners at different proficiency levels failed to notice the distinction between the literal meanings and implied meanings and 2) the degree to which learners at different proficiency levels may have responded to a pragmatic meaning that was at a level beyond intended meaning.

6.6 Implications for test development

Analyzing learners' ability to comprehend conversational implicature was the focus of this study due to it being widely understood to be a central part of listening, and because it is already included in many language comprehension tests, although whether there is a separate implicature ability that can be isolated has been unclear to date. In light of this study's results, a key and obvious question would seem to be whether or not it is

important to test conversational implicature in a targeted manner. It is the author's view that it depends on the context and construct in question. That is, it requires the construction and consideration of one's validity argument. Returning to Kane's argument-based approach to validity discussed in Chapter 3, below is a revisiting of the specific links in the chain of inferences that were of interest—when one assumes that comprehension of conversational implicature is an important component of the construct.

6.6.1 Validity argument

In Chapter 3 it was asserted that it is rare to see validity arguments made in conjunction with actual test content, and this study was framed as a validity investigation for a possible subskill of listening. Specifically, three of the inferences in Kane's chain of inference validity approach were highlighted as being inferences for which evidence was being sought (Figure 31 below).

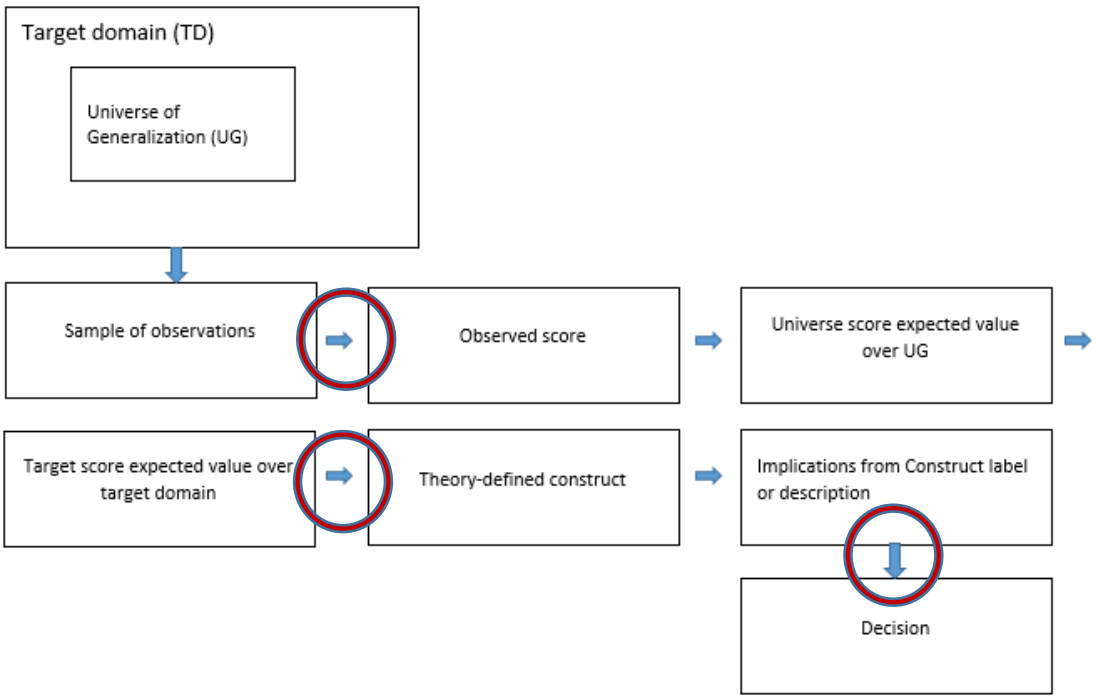


Figure 31. Interpretive Argument Chain

Working within the target domain of "interacting competently in conversation" and trying to make predictions about the universe of generalization that can be called "understanding conversational implicature," the context of this study narrows to performance on listening implicature items as being a sample of observations for understanding conversational implicature. That leads to the first link of interest in the chain being the need to provide evidence from the *Sample of observations* to *Observed score*. Evidence for this inference is provided here via the observation that scores from CR format converge with MC format, as was discussed in detail in section 5.3.1. The results of this study also showed, via Table 10 in section 5.3.1, that target-level learners who were able to obtain an observed "correct" score by selecting the key provided to them in MC format could also obtain the key when not provided with it, demonstrating the necessary and desired understanding, and therefore evidence of the meaningfulness of

the observed scores. The adequate reliability of both format's measures further supports the inference of observed scores from the TCI, whether in MC format or CR format, as being meaningful. It also provides some evidence for the inference from *Universe expected value over UG* to *Target score expected value over target domain*. This evidence comes from the correlation between the ability to generate a correct understanding of a conversational implicature with a selected response task to the ability to do so on productive tasks. This correlation can be interpreted as support for the notion that in the target domain of interacting in conversations that entail implicature, there is a better chance of success for those that did well on the TCI than for those that did not.

The second circled inference, which was identified as a focus from the study's start, relates to the idea that understanding of the theoretical construct would be improved by providing evidence that implicature and general items, while both part of listening construct, are identifiably distinct within that larger construct. As already discussed in the "Research Question 2" section in this chapter above, the results of this study's analyses point to it being unlikely that there is an identifiable, stable "listening for implicature" skill within listening proficiency. These results therefore support a theory-defined construct of "listening for implicature" that is part of a broader listening proficiency construct, and that for test developers, this means not always needing to use comprehension of implicature items in listening tests. This leads naturally to the final chain of inferences in Kane's argument, that test developers have evidence to support test score use, or the decisions made on the basis of test scores. Unidimensionality of listening would give support for making decisions about expectations of a listener's implicature ability on the basis of their general ability scores alone: those with advanced

level general listening will likely be able to understand implicatures; those at intermediate level will be likely be able to regularly understand implicatures, and those at lower proficiency levels will be able to do so with limited frequency.

In summary, we see here, that by using Kane's argument-based approach for validity, which allows test developers to lay out step by step the way evidence may or may not support their claims, that if there is a need for assessing conversational implicature in a testing context, doing so through multiple-choice format items is likely a defensible approach—but, because of the evidence for unidimensionality of listening, in a number of assessment contexts, distinct implicature items are not likely necessary to obtain evidence of interactional competence for listeners.

6.6.2 CEFR

Consideration of this study's results in light of the CEFR was also one of its aims. The widely used CEFR posits that within the receptive skills of listening and reading, "implicit" understanding across a range of language-use contexts is an expected, rather than a developing, part of the profile at C1. The results from this study appear-to give credence to that expectation within the language-use context of listening to proficient speakers engage in naturalistic conversation. The participants who were identified as C1 by the independent measure of proficiency (the Michigan English Test) performed in marked contrast to the B1 and B2 participants on the implicature items. They also performed markedly better on the general items, but the fact that the implicature items mirrored this pattern (box plots from section 5.5 replicated in Figure 32 below) lends support to the idea that while lower-level proficiency listeners might be able to process and understand implicature, it is only when they have reached higher overall proficiency

that that they will be able to do so with regularity—or potentially the "ease" that the CEFR's "Understanding Conversation Between Other Speakers" scale C1 level descriptor asserts.

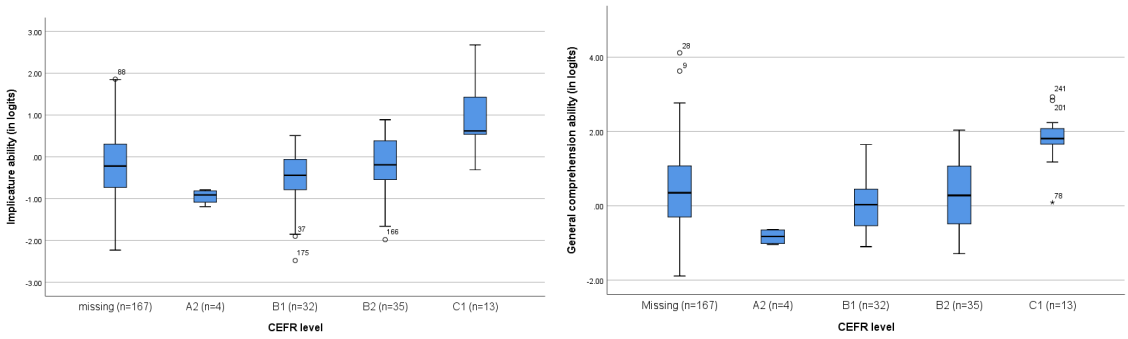


Figure 32. Participants by CEFR level for Implicature (left) and General (right) Ability

It is also worthwhile to consider these results in regard to the question of the CEFR being underspecified. As was discussed in Chapter 3, the CEFR was originally created as a point of reference and not as a specifications document. Critiques of under-specification are only applicable if one is trying to use the framework as a specification document. In the case of this study, the item specifications were developed without consideration of the CEFR. Rather, they were drafted with the goal of specifying listening items that could assess real-time comprehension of natural conversational exchanges between highly proficient users of the language—when they were employing conversational implicature and when they were not. The analyses of test-takers by CEFR level that were done then actually serve to support the use of the CEFR as originally intended—what we see is that language users with a listening comprehension ability of at least C1 level are likely going to be able to interpret conversational implicature when deployed by two proficient speakers conversing with each other without making any adjustments in their language for speed or structure.

6.7 Limitations

A number of the study's limitations need to be noted. One limitation was not employing a linking item design (i.e., using the same 8 or 10 items across all seven versions of the measures). Using overlap of different groups of items across the seven forms did create the linkage needed for the Winsteps program to make Rasch difficulty estimates, but it did not allow for strong enough linkage for any estimates on items that participants did not sit for. A 50-item measure with common linking items across all would have also increased confidence in the CFA results as there would have been more linkage across the indicator variables that were used in those models. Also related to scoring, while the CR inter-rater reliability was adequate, the inclusion of a consensus or discussion step of discrepant rating would have likely improved that reliability.

As noted in discussion of the short-term memory results, unexpected behavior of the participants on the forward auditory digit span task made those results unusable, and a failure by many participants to follow the directions for the backward auditory digit span made those results questionable. This prevented the exploration of the role of short-term memory on participant performance or the relationship of short-term memory to working memory in a meaningful way.

Another limitation was the absence of an introspection protocol step with a subset of participants to gather qualitative information about impressions of the two different item formats and to lend support to item developers' judgments about accessibility of topics and implicatures. This step was one that was recommended during the study setup but the logistics of administration prevented it from being feasible. Such a step would greatly enhance assumptions made about the comparability of constructed response and

multiple choice items, as well as potentially provide insights on particularly difficult or easy implicature items. With the items used in this study being made available for further research, hopefully this limitation can be remedied in future.

Another potential limitation was the proficiency level of participants. The assertions made about the unidimensionality of the measure used in this study would be even stronger if a roughly equal percentage of B1, B2, and C1 level participants were in the sample, as was the goal by design, rather than the preponderance of B1 and B2 level participants and a smaller group of C1 level. A follow-up study that involves a distractor analysis of MC items by participants in the three different CEFR groupings may shed light on this.

6.8 Conclusion

The results of this study support a conclusion that listening cannot be subdivided into a comprehension of implicature trait and a general listening trait. All subcomponents of listening were not looked at, but by not being able to separate these two "subskills," there is evidence that listening is unidimensional. Whether listening for conversational implicature is inherently a higher-level proficiency skill is still an open question. What can be concluded from this study, however, is that when conversational implicature is embedded in natural dialogues and must be processed in one listening—as is often the case in real-life communication—intermediate listeners are only moderately capable of comprehension whereas higher level listeners (defined as C1 on the CEFR) are capable of comprehension. Further, the results of WM failing to correlate with implicature listening ability, while correlating weakly with general listening ability, which is in line with past correlations found between WM and language ability, seem to indicate that if there is a

separate ability for comprehending conversational implicature, it may be linked to non-linguistic ability.

In terms of test development practice, while it was demonstrated that evidence was gathered to support an interpretive argument for the use of conversational implicature items in a general listening test, if implicature is not a central part of the test purpose (e.g., testing comprehension for learners who will be using their language ability in service encounters), focusing on general listening will likely produce a result that can then allow inferences to be made about the broader or more general skill of listening. Replication of this assertion is welcome and for this reason a 52-item version (26 general items and 26 implicature items) of the 60-item test used in this study will be made available for other researchers to use with different samples of English language learners and to analyze in contrast to other measures. Furthermore, this instrument could be used to extend this study, as was discussed above, both by administering items with targeted versus more-open stems and with two different types of intonation to further explore the salience of intended meaning versus pragmatic stance of speakers. As conversational implicature is such an ever-present aspect of communication and comprehending those implicatures is such a necessary skill, the more test developers and applied linguists understand it and how it relates to other types of listening, the better able we will be to create listening measures that generate valid listening scores.

Appendix A: Listening test items

LI_01
Domain: personal; Topic: travel/vacation
Turns: 2; Words: 29
Implicature type: relevance (conversational implicature)
Degree of implicature: low (general context not relevant, specific context provided, number of cues to interpret 2 - car got hit, wallet was stolen)

W: How was your trip to New York?

M: First my car got hit when I left it parked overnight on the street, and then my wallet was stolen.

How does the man feel about his trip to New York?

It wasn't enjoyable.*
It wasn't long enough.
It was boring
It was expensive.

LI_02
Domain: Occupational; Topic: workplace situation (take a break)
Turns: 4; Words: 52
Implicature type: relevance (conversational implicature)
Degree of implicature: medium (general context not provided, specific context not provided, number of cues to interpret: 1, a ton of work to do)

M: Wow, it's really beautiful outside. Not a great day to be cooped up inside.

W: Yeah, it is. Let's go grab a cup of coffee, take a walk around the block or something.

M: Sounds good, but I've got a ton of work to do.

W: Suit yourself. I'm going to go stretch my legs.

What will the man do?

Continue working*
Buy the woman a coffee
Stretch his legs
Go outside later

LI_03
Domain: Occupational; Topic: student behavior
Turns: 4; Words: 46
Implicature type: relevance (avoidance)
Degree of implicature: high (general context not provided, topic not provided)

M: Students aren't allowed to go to the bars on South Street, are they?

W: Right. They're off limits to all students.

M: I see. But, how often do students [emphasize "try"] actually try to go?

W: Those bars are off limits to the students.

Why does the woman repeat her response?

She doesn't want to answer the question.*
She didn't understand the question.
She thinks the man didn't hear her.
She thinks the man doesn't know where they are.

LI_04
Domain: Occupational; Topic: workplace situation
Turns: 4; Words: 49
Implicature type: indirect criticism (conversational implicature -- this type *can* be conventional if in response to enquiry "What do you think of the handbook?" "It looks like it was printed in 1990."
Degree of implicature: low (general context not provided, specific context is provided, number of cues to interpret 1 - looks like printed in 1990)

M: Sorry to bother you, Karen, but I have a quick question: Is this the current staff handbook?

W: The one with the yellow cover? Yeah, that's it.

M: This looks like it was printed in 1990.

W: Maybe not quite that far back, but it does need updating.

What does the man say about the handbook?

It is old.*
It is confusing.
It looks ugly.
It was difficult to find.

LI_05
Domain: occupational; Topic: number of staff
Turns: 4; Words: 44
Implicature type: indirect negative evaluation
Degree of implicature: medium

M: How many people are on your team now?

W: Well, after the layoffs, we're down to three --including myself.

M: Really? And you guys were struggling to keep up before, weren't you?

W: Yeah, we'd be fully staffed with six people. With five, it was tough, but we were getting by.

How does the woman feel about her team?

She doesn't have enough employees.*
They don't work hard enough.
They work well together.
She likes most of her employees.

LI_06
Domain: Educational; Topic: Volunteering
Turns: 3; Words: 76
Implicature type: relevance (avoidance) -- borderline conventional
Degree of implicature: high (general context not provided, specific context is provided, number of cues to interpret 6 - all the emails; still need a student; I'm taking four classes; part-time job has gone from ten hrs to twenty; only once a month; look good on CV)

M: Janet, you've probably seen all the emails in the last week … we <u>still</u> need a student from our program to volunteer to represent us in the college senate.

W: Yeah, Professor, I know. But I'm taking four classes this semester AND my part-time job has gone from ten hours a week to twenty.

M: Well, the senate does meet only once a month, so it's only a three- or four-hour-per-month commitment. And it'll look good on your C.V.

Why does the woman mention the number of courses she's taking?

Because she doesn't want to participate in the college senate*
Because she wants the man's help
To explain why she didn't respond to the man's email message
To explain why she wants to reduce her work schedule

LI_07
Domain: Educational; Topic: Group study
Turns: 3; Words: 49
Implicature type: indirect criticism
Degree of implicature: low (general context not provided, specific context is provided, number of cues to interpret 1 - she doesn't strike me as a group study kind of person)

W: Don't you think it'd be a good idea to ask Laura to join our study group? She's easily the smartest person in the class.

M: Sure, ask her if you want. But she doesn't strike me as a "group study" kind of person.

W: Maybe. But there's no harm in asking

What does the man imply about Laura?

She prefers working alone.*
She already refused their request.
She is not as smart as the woman thinks.
She is in a different study group.

LI_08
Domain: Personal; Topic: recreational activities
Turns: 5; Words: 57
Implicature type: irony
Degree of implicature: high (general and specific context not provided, number of cues to interpret 3: performance on river; thought you were an experienced kayaker; going backwards .. intentional?)

W: Jack, that was quite a performance on the river.

M: What do you mean?

W: I thought you said you were an experienced kayaker?

M: I am, sort of. I mean, I've gone kayaking about a dozen times in the last couple of years.

W: Oh. So when you were going backwards for about two hundred meters there, that was intentional?

What does the woman imply about Jack?

He isn't skilled at kayaking.*
He is a better kayaker than she is.
He finished quicker than she expected.
He shouldn't be allowed to kayak.

LI_09
Domain: Occupational (personal); Topic: Objects (bike)
Turns: 4; Words: 54
Implicature type: relevance (easy)
Degree of implicature: medium (general context not provided, specific context is provided, number of cues to interpret: 4, chaining it up; not a strong deterrent; local criminal element; that's terrible; less than five weeks of use)

W: Did I see you driving in this morning, Kevin? I thought you bought a bike to commute.

M: I did. But apparently chaining it up at the bike rack in my apartment complex wasn't a strong enough deterrent for our local criminal element.

W: Oh, that's terrible.

M: Yep, $400 for less than five weeks of use.

What does the man say about his bike?

It was stolen.*
It needs repairs.
He has no place to lock it.
He wants to sell it.

LI_10
Domain: Occupational; Topic: workplace situation
Turns: 3; Words: 34
Implicature type: irony
Degree of implicature: medium (general and specific context provided, number of cues to interpret: 3, standing around awkwardly; trying to make small talk + tone)

M: We're having a cake in the breakroom to celebrate Sam's wedding. Are you coming?

W: [*without enthusiasm*] Sure. Who doesn't enjoy standing around awkwardly trying to make small talk that isn't work-related.

M: Exactly! But delicious, free cake…

What does the woman imply about the event?

It won't be fun.*
The cake will not be good.
People will only talk about work.
She doesn't have time to go.

LI_11
Domain: Personal; Topic: everyday life (dentist)
Turns: 4; Words: 28
Implicature type: indirect negative evaluation
Degree of implicature: low (general context not relevant, specific context provided, number of cues to interpret: 1, I wouldn't go that far)

W: How long have you been going to your current dentist?

M: Forever. More than ten years, I guess.

W: Oh, so he's pretty good.

M: Uh, I wouldn't go that far.

What does the man say about his dentist?

He is good enough.*
He is located too far away.
He is the best one in the area.
He is not experienced.

LI_12
Domain: Educational; Topic: attend demonstration
Turns: 2; Words: 34
Implicature type: indirect criticism
Degree of implicature: medium (general context not provided, specific context is provided, number of cues to interpret: 1, they're gonna do what they're gonna do)

M: Are you coming to the rally on Sunday? It's to protest the plan to knock down Ferber Library.

W: I don't know. It's a nice idea, but ultimately, they're gonna do what they're gonna do.

What does the woman say about the rally?

She doesn't think it will be effective.*
She won't have time to attend it.
She doesn't understand its purpose.
She disagrees with the people organizing it.

LI_13
Domain: Educational; Topic: Academic situation (professor feedback)
Turns: 5; Words: 42
Implicature type: indirect criticism
Degree of implicature: medium (general and specific context provided, number of cues to interpret: 3, two comments on twenty-page paper; took me five weeks to write; no the point; more than one implicature available)

M: Professor Haley made two comments on my final paper.

W: What grade did you get?

M: An A, but just two comments on a twenty-page paper that took me five weeks to write?

W: But, you got an A.

M: That's not the point.

What is the man implying about his professor?

He didn't read the paper carefully.*
He didn't explain the assignment well.
He doesn't grade students' fairly.
He doesn't give students time to finish their work.

LI_14
Domain: Personal/Public; Topic: building
Turns: 3; Words: 48
Implicature type: irony
Degree of implicature: medium (general context not relevant, specific context is provided, number of cues to interpret: 3, thrilled about the fact, spent half a million dollars + tone)

M: So this is the new city hall building? This wasn't here last time I visited, was it?

W: No, it's brand new. And we're all just [*sarcastic*] thrilled about the fact that they spent half-a-million dollars on that fountain in the courtyard.

M: Wow. And it's not even that cool looking.

How does the woman feel about the fountain?

It was too expensive.*
It wasn't approved by the city.
It needs better maintenance.
It isn't well designed.

LI_15
Domain: Occupational; Topic: workplace situation (data)
Turns: 4; Words: 67
Implicature type: indirect criticisim/negative evaluation
Degree of implicature: high (general context not provided, specific context is provided, number of cues to interpret: 4, can't believe way consultants submitted data; take an hour to spell out what's wrong; take them another couple weeks to fix + tone)

M: I can't believe the way the consultants submitted the data. Information repeated in different places with different labels, inconsistent coding…

W: So, send it back. We're paying them!

M: I know. But it'll take me at least an hour to spell out everything that's wrong with it and <u>then</u> it'll take them another couple weeks to get it fixed…

W: Oh, no. I think I know where this is going.

Why is the woman worried?

The man will suggest they fix the problem themselves.*
The man will ask her to visit the consultant.
The man wants to hire extra staff for a task.
The man wants to send the data to the consultant.

LI_16
Domain: Occupational; Topic: workplace situation
Turns: 2; Words: 39
Implicature type: indirect criticism
Degree of implicature: low (general and specific context provided, number of cues to interpret: 2, response in 24-48 hours; getting something resolved is different story)

M: When you submit a request to the tech department for help with something, how long does it usually take them to respond?

W: You'll get a response in twenty-four to forty-eight hours. Getting something resolved is a different story though.

What does the woman imply about the tech department?

They do not solve problems quickly.*
They do not respond to questions sent by email.
They are taking longer than usual to respond.
They will fix the problem the following day.

LI_17

Domain: Educational; Topic: academic situation (quiz)
Turns: 2; Words: 34
Implicature type: relevance
Degree of implicature: low (general and specific context provided, number of cues to interpret: 1, we'll be grateful for all these quizzes)

M: Did Professor Smith say we have another quiz tomorrow?

W: Yeah. The syllabus says we have one every week. But I bet we'll be grateful for all these quizzes by the end of the semester.

What does the woman think about the quizzes?

They will help the students learn.*
They are more difficult than the exam.
They occur too frequently.
They should be listed on the syllabus.

LI_18
Domain: Occupational; Topic: workplace situation (supervisor)
Turns: 4; Words: 36
Implicature type: relevance (avoidance)
Degree of implicature: High (general context not provided, specific context is provided, number of cues to interpret: 4, umm; tough question; don't have strong opinion), more than one implicature available

W2: Do you think George is a good boss?

M1: Umm. [*hesitantly*] That's a tough question.

W2: I mean, what do you think about the way he talks to the employees?

M1: I really don't have a strong opinion about that.

Why didn't the man answer the woman's questions?

He doesn't want to talk about the boss.*
He doesn't know who her boss is.
He doesn't understand her.
He doesn't have time.

LI_19
Domain: Educational; Topic: recreational activities (sports)
Turns: 3; Words: 33
Implicature type: indirect criticism
Degree of implicature: medium (general context not relevant, specific context is provided, number of cues to interpret: 3, how much is it?; have you ever even gone?, + tone)

W: Do you wanna go to the football game this weekend?

M: I don't know. How much is it?

W: [*incredulous]* How much is it? Have you ever <u>even</u> gone to a game? It's <u>free</u> for students.

Why does the woman repeat the man's question?

She's surprised he doesn't know tickets are free.*
She's upset he didn't already buy tickets.
She thinks more students should attend football games.
She wanted to show that she heard him.
Alt: She was unable to hear what he said.

LI_20
Domain: Educational; Topic: academic situation (research)
Turns: 2; Words: 42
Implicature type: irony (conversational implicature)
Degree of implicature: low (general context not relevant, specific context provided, number of cues to interpret: 2, I haven't heard much from my colleagues recently; out of sight out of mind)

M: How's it going with your research project? Have you been able to keep making progress despite being away from the university?

W: [*slightly sarcastic*] Oh yeah, the one thousand mile distance from my colleagues *and* the time zone difference have had *really* no impact at all…

What does the woman mean?

She is not staying in touch with her colleagues.*
She will return to her university soon.
She is too busy to finish a project.
She wants to change her research project.

LI_21
Domain: Public; Topic: everyday life (shopping)
Turns: 4; Words: 38
Implicature type: relevance
Degree of implicature: low (general context (grocery store) not provided but specific context (shopping) is provided, number of cues to interpret: 2, repetition of 7.99; I remember when they used to be 79 cents)

W: Excuse me, how much are the cherries?

M: Just seven-ninety-nine a pound, ma'am. And just picked yesterday.

W: Seven-ninety-nine... I remember when they used to be seventy-nine cents a pound.

M: Sorry to say, ma'am, but those days are long gone.

What does the woman imply about the cherries?

They are expensive.*
They are several days old.
They will not sell well.
They don't look delicious.

LI_22
Domain: Personal; Topic: everyday life (apartment)
Turns: 2; Words: 21
Implicature type: indirect criticism
Degree of implicature: high (general context and specific context not provided, number of cues to interpret: 1) multiple implicatures available

M: I heard you moved downtown. How are you liking your new apartment?

W: The rent is lower; I can say that much.

What does the woman imply about her apartment?

She doesn't like much about it.*
She can't afford the rent.
She likes the neighborhood
She thinks the man would like it.
Alt: She needs a roommate

LI_23
Domain: Educational; Topic: academic situation (conference)
Turns: 4; Words: 58
Implicature type: irony
Degree of implicature: medium (general context not provided, specific context provided, number of cues to interpret: 3, waiting for visa; booked everything; possibility of being out a lot of money)

M: I'm really looking forward to the conference in Toronto. You're going, right?

W: If my visa comes through in time.

M: The conference is next week. Are you still waiting to book your flight and hotel and all that?

W: No, I've booked everything. And it'll really be awesome to have spent five hundred dollars on a conference I don't even go to!

What is the woman's concern?

She won't be able to attend the conference.*
She can't afford the conference.
She couldn't find a hotel near the conference.
She doesn't want to attend the conference alone.

LI_24
Domain: Personal; Topic: daily life (household task)
Turns: 2; Words: 28
Implicature type: relevance
Degree of implicature: high (general context not provided, specific context is provided, number of cues to interpret: 3, going to rain all day; was planning on painting garage; tomorrow's forecast not a cloud in sky)

M: Is it really going to rain all day? I was planning on painting the garage today.

W: Well, tomorrow's forecast says there won't be a cloud in the sky.

Why does the woman mention tomorrow's weather?

To suggest when the man can do the work*
To explain why she is changing her plans
To remind the man how bad the weather will be
To support her decision to stay at home

LG_25
Domain: Personal; Topic: vacation / recreational activities
Turns: 5; Words: 70
Subskill: General understanding

M: You've been out to Willow Island before, haven't you?

W: Oh, yeah. It's beautiful. I've been there a few times.

M: What's your recommendation for covering it all in a day? The website says there are boat tours, bike rentals -- there's even a horseback option.

W: My favorite thing is to hike it -- but I've always been there for a couple of nights.

M: Wish we could do that, but it's just not possible.

What are the speakers discussing?

How to see a particular location*
A vacation they took together
Their favorite outdoor activities
Transportation choices in the city

LG_26
Domain: occupational; Topic: everyday life (driver's license)
Turns: 4; Words: 64
Subskill: General understanding

W: Were you on a sales call, Frank? I haven't seen you at your desk all day.

M: No. I went to the department of motor vehicles to renew my driver's license. I was there for two hours.

W: Oh. Yeah, motor vehicles is never fun. But at least you got it done.

M: No, that's the worst part. I forgot a form. I have to go back tomorrow.

What are the speakers discussing?

What the man was doing earlier*
Where to obtain a driver's license
Why the man missed a meeting
Where the man is getting his car fixed

LG_27
Domain: educational; Topic: academic situation (discussing a class)
Turns: 2; Word count: 46
Subskill: General understanding

M: How's your stats class going?

W: Surprisingly well. I thought I would hate it because I'm terrible with numbers. But the professor is great. He makes the material really interesting-- and knowing what I'm learning is gonna be super-useful when I'm looking for a job, I'm sure.

Why is the woman surprised?

She enjoys her statistics class.*
She did well on an exam.
Her professor wants her to study statistics.
Her professor offered her a job

LG_28
Domain: Personal; Topic: sports
Turns: 4; Word count: 45
Subskill: General understanding

M: Did you watch the baseball game last night?

W: I did-- what an ending!

M: Yeah. O'Brien is the last guy in the world I would pick for pitching the final inning, but boy did he come through.

W: I know-- I think he's my new favorite player!

What do the speakers agree about?

The game was exciting.*
The game took too long.
Watching games together is fun.
O'Brien is the team's best player.

LG_29
Domain: personal; Topic: everyday life (friend's diet/exercise regimen)
Turns: 3; Word count: 46
Subskill: General understanding

W: Tom is really taking his new exercise routine seriously-- talk about a change!

M: Yeah, tell me about it. The last time I saw him I almost didn't recognize him.

W: Yeah, he looks good. And maybe we'll even be able to talk him into joining the softball team.

What are the speakers discussing?

A friend's appearance*
A friend's illness
How often their friend exercises
Their softball team

LG_30
Domain: Educational; Topic: recreational activities
Turns: 3; Words: 51
Subskill: General understanding

M: Did you hear which band is being invited to this year's Autumn Festival?

W: Yeah, everyone's talking about it. The Winslow Brothers? Have they even made an album since we were born?

M: I doubt it. The faculty's going to have a lot more fun at that show than any of us.

Why are the students upset?

They don't like the entertainment for an event.*
Tickets to an event are sold out.
The band didn't play the songs they expected.
Their favorite band isn't coming to their school.

LG_31
Domain: Educational; Topic: employment
Turns: 5; Words: 51
Subskill: General understanding

M: Did you see the emails about the job fair? It sounds like it's going to be huge.

W: I know. They're saying there'll be like 80 companies there.

M: Better print out a bunch of resumes!

W: Unfortunately I won't be able to make it. I have two tests on Wednesday.

M: That's too bad.

What are the speakers discussing?

An opportunity to meet employers*
A job interview the woman had
How to write a resume
The woman's exams

LG_32
Domain: Educational; Topic: academic situation (homework)
Turns: 3; Words: 48
Subskill: General understanding

M: I don't think I've ever taken a class where the professor gave homework grades more weight than test grades.

W: I think it makes sense. Homework shows that you can apply what you know. Tests just show what you memorized.

M: Yeah, but people can help you with your homework!

What are the students discussing?

Their teacher's grading policy*
Their grades on a homework assignment
Doing a homework assignment together
How difficult their test was

LG_33
Domain: Personal; Topic: daily life (cost of living)
Turns: 4; Words: 43
Subskill: General understanding

M: The price of gas here is so expensive.

W: It's not just gas. Food, rent—everything.

M: Yeah, I'm starting to notice that. I mean, it's a great city, but I don't think I could ever live here.

W: Well, I have gotten used to it.

N: What are the speakers discussing?

How much it costs to live in the city*
How much they enjoy living in the city
How cheap the man's apartment is
How to use the transportation system

LG_34
Domain: Public; Topic: lost object
Turns: 3; Words: 53
Subskill: General understanding

M: Excuse me, I was in here yesterday afternoon for lunch and I think I left my sunglasses behind. Do you know if anyone turned them in?

W: Well, uh, you can check for yourself. Let me go get the lost-and-found box—there are actually probably two dozen pairs of sunglasses in there.

M: Great, thanks.

N: Why is the man talking to the woman?

He lost something.*
He is looking for someone.
He wants to make a reservation.
He found some sunglasses at his table.

LG_35
Domain: Occupational; Topic: rescheduling a meeting
Turns: 4; Words: 64
Subskill: General understanding

W: Jerry, I see that you have the Walker Conference Room booked on Friday at two o'clock. Could you reschedule that meeting?

M: Probably... It's just my weekly team meeting. But why?

W: The vice president from Orion Corporation will be here that afternoon. And as you know, the Walker Room is the nicest one we have, so I'd like to use it.

M: Sure, no problem then.

N: Why does the woman need the conference room on Friday?

An important person is visiting the office.*
She needs to hold a team meeting.
She wants to discuss something private with the man.
Several clients will be in the office at the same time.

LG_36
Domain: Personal; Topic: favor
Turns: 4; Words: 62
Subskill: General understanding

M: Margaret, I have a favor to ask.

W: [*cautiously*] Okay…

M: I have to go out of town for almost a month, and my plants won't survive without water… My neighbor used to do it, but she moved away, and I have no one else to ask…

W: No problem, I go by your place on my way to work anyway, so no trouble at all.

N: Why does the man need the woman's help?

He will be traveling for several weeks.*
He is moving into a new apartment.
He is having problems with his neighbor.
He has no way to get to work.

LG_37
Domain: Educational; Topic: academic situation (professor)
Turns: 4; Words: 68
Subskill: General understanding

M: Did you see that confusing email from Professor Williams?

W: No, what's it about?

M: It talks about our next assignment, but it totally contradicts what he said in class on Tuesday—and neither the email or the stuff he said in class match what's in the syllabus.

W: [*slight exasperation*] Huh, so after only two weeks we're learning why everyone says you need to double-check, and then triple-check everything for his class.

N: What are the speakers saying about their professor?

He isn't well organized.*
He cancels too many classes.
He doesn't respond to email.
He didn't give them a syllabus.

LG_38
Domain: Educational; Topic: proofreading freelance
Turns: 3; Words: 63
Subskill: General understanding

M: Could you recommend one of your graduate students for a proofreading job?

W: I think so—but what kind of proofreading exactly?

M: I'm submitting a thirty-page paper for publication, and I'd like someone to give it a really close read. Grammar isn't one of my strengths, so I'm looking for someone who's really accurate with language. A deep understanding of the topic isn't necessary.

N: What does the man need?

Someone to review his writing*
Advice on an idea for a project
A letter of recommendation
Help with his research

LG_39
Domain: Occupational; Topic: workplace security
Turns: 4; Words: 52
Subskill: General understanding (significant detail)

W: You'd better hope that Fred doesn't catch you doing that.

M: Doing what?

W: Propping the door open like that. In case you didn't know, he's a real stickler for security. Open doors, among other things, drive him crazy.

M: Yeah, I do know. But I've got a bunch of boxes to move back here.

N: What has the man explained?

Why he is leaving a door open*
Why he needs help moving the boxes
Why he is concerned about security
Why he didn't talk to the security guard

LG_40
Domain: Occupational; Topic: workplace situation
Turns: 4; Words: 60
Subskill: General understanding

M: Have things gotten any quieter for you at work?

W: Not really. I met a big project deadline this past Monday, which felt good-- but I've already been put on two other projects.

M: Sounds like you need a vacation.

W: I wish. But if I go on vacation, it just means a bunch of stuff won't get done till I come back.

N: What is the woman complaining about?

She has too much to do at work.*
Her boss changed a deadline.
She had to cancel her vacation.
Her staff doesn't work hard enough.

LG_41
Domain: Educational; Topic: driving on campus
Turns: 4; Words: 48
Subskill: General understanding

M: I think you ought to slow down.

W: Really? I'm only driving twenty miles an hour…

M: Yeah, but the official speed limit on campus is ten miles an hour. Which makes sense, considering that there are hundreds of students walking every-which-way all day long.

W: Okay, okay, I get it.

N: What did the man explain to the woman?

Why she should drive slowly*
Where she should park
Which direction to go
Why the campus is crowded

LG_42
Domain: Personal; Topic: objects (bookcase)
Turns: 3; Words: 39
Subskill: General understanding

M: That's a nice bookcase. Is it new?

W: Actually, I made it myself. I took a carpentry course last summer at the community center.

M: Wow. That was money well spent. I should check out one of those courses for myself.

N: What does the man think about the bookcase?

It is well made.*
It is valuable.
It is nicer than his.
It is very modern.

LG_43
Domain: Educational; Topic: academic situation (number of classes)
Turns: 3; Words: 53
Subskill: significant detail

W: You look exhausted, Kevin. Are you getting enough sleep?

M: Not really. This semester is crazy—I'm taking five classes, and four of them involve really demanding, semester-long projects.

W: Sounds challenging. But don't forget that the good thing about being in university is that the semester eventually ends! Wait till you're working full time.

N: Why is the man so busy?

He is taking several difficult classes.*
He is studying for his final exams.
He is looking for a new job.
He is working and studying.

LG_44
Domain: Public; Topic: Exercise/fitness
Turns: 4; Words: 64
Subskill: General understanding

M: Excuse me, you work here at the gym, don't you?

W: Yes, I'm a trainer. How can I help you?

M: Well, I've been coming here to exercise for six months now, pretty regularly. But I'm not seeing any results. So, I think I could probably use some professional advice.

W: That's what we're here for. Let me check the calendar, and schedule you for a session.

N: What does the man want to do?

Improve his physical fitness*
Apply for a job as a personal trainer
Learn what the gym's rules are
Purchase a gym membership

LG_45
Domain: Public; Topic: hotel stay
Turns: 4; Words: 61
Subskill: General understanding (significant detail)

M: Hi, I just checked in five minutes ago, into room one-oh-seven.

W: Yes, hi. Is there a problem?

M: Yeah. The air conditioning unit is leaking pretty heavily onto the floor. If you could switch me to another room, that'd be great.

W: Oh, sure, that shouldn't be a problem, sir. And sorry about that. Just give me a minute to check what's available.

N: What is the man requesting?

To move to another room*
To have his room cleaned
To change his reservation
To have the air conditioner fixed

LG_46
Domain: Occupational; Topic: Workplace situation
Turns: 4; Words: 53
Subskill: General understanding

W: So you said this morning that you finally finished the report on the Thompson Project? Must be a good feeling.

M: Actually, I spoke too soon. Mary's asked me to do some revisions.

W: Really? Did you leave something out?

M: No, it's more a matter of tone. She said it was a bit too negative.

N: What are the speakers discussing?

A report the man wrote*
An assignment that is late
Their coworker's personality
The results of their project

LG_47
Domain: Educational; Topic: academic situation
Turns: 3; Words: 52
Subskill: General understanding

W: Tom, are you willing to admit that you didn't put your best effort into this paper?

M: Yeah, probably not. I left it until the last minute.

W: Okay. Well, I'm going to give you a week to revise it, because otherwise this grade is going to wreck your grade for the entire semester.

N: What does the man have to do?

Work on an assignment again*
Pay more attention in class
Stop arriving late to class
Meet with a tutor

LG_48
Domain: Occupational; Topic: animals
Turns: 4; Words: 60
Subskill: General understanding

W: Did you hear that crazy story about a bear showing up in the parking lot yesterday evening?

M: A bear in the parking lot? Are you serious?

W: Yeah, it was on the ten o'clock news, and it's all that anyone else around the office is talking about.

M: Wow. That is crazy. I hope they caught him and took him far away…

N: What do the speakers think is surprising?

A wild animal appeared in the parking lot.*
A bear was in the parking lot and nobody noticed.
The parking lot will be closed for several days.
A television news team is in the parking lot.

LG_49
Domain: Occupational; Topic: job application
Turns: 4; Words: 55
Subskill: General understanding

M: You look kind of gloomy, Kara. What's wrong?

W: I just found out that I'm not getting the position in the marketing department that I applied for.

M: Oh, really? I didn't know that you were interested in marketing.

W: [*sighing*] Yeah, I am, but I don't have any experience. Which I guess explains why I was passed over.

N: What happened to the woman?

She wasn't chosen for a job she wanted.*
She has to take a class she's not interested in.
She is being transferred to a different department.
She was criticized by the director of marketing.

LG_50
Domain: Personal; Topic: sports
Turns: 4; Words: 56
Subskill: General understanding/significant detail

W: Did you watch the game last night?

M: I started to, but I fell asleep during the second half.

W: You fell asleep! But it was the championship, and the most exciting game in at least a decade.

M: I know, I know. But I was up till two a.m. the night before finishing up a report for work.

N: Why is the woman surprised?

The man didn't watch the entire game.*
The man isn't interested in sports.
The man didn't go to the game.
The man didn't know the game was on TV.

LG_51
Domain: Public; Topic: everyday life (children behavior)
Turns: 4; Words: 64
Subskill: General understanding

M: Here's your check, ma'am. I hope you all enjoyed your meal.

W: Thank you, we did.

M: And I have to say, ma'am, your children are really well-behaved. Such sweet kids! Most of the kids who come into the restaurant can't keep still.

W: Oh, thank you. But I think today was just your lucky day. Ordinarily they're climbing all over everything like all the other kids.

N: What are the speakers discussing?

The behavior of the woman's children*
Why the family came to the restaurant
Activities that children enjoy
How many families were in the restaurant

LG_52
Domain: Occupational**;** Topic: workplace situation (late for meeting)
Turns: 3; Words: 64
Subskill: General understanding (significant detail)

W: Why are you so late? You know that the clients were here at eight a.m. waiting for you, right?

M: [*flustered*] Yes, yes, I know. I actually took a taxi this morning instead of my usual bus to get here earlier, but first the driver got lost, and then he got a flat tire.

W: Okay. Well, James adjusted the schedule, so you're presenting at ten-thirty now.

N: Why was the man late?

He had problems with a taxi.*
He missed his usual bus.
He forgot about an important meeting.
He didn't have the correct schedule.

LG_53
Domain: occupational**;** Topic: workplace situation (clothing)
Turns: 3; Words: 42
Subskill: General understanding

W: You look nice, Stanley. I don't think I've ever seen you wearing a sport coat in the office before.

M: I'm on the hiring committee for the new office manager. We're doing four interviews today.

W: Oh right. Need to look professional for those.

N: Why is the man dressed differently?

He has several important meetings.*
He is interviewing for a new job.
His office hired a new manager.
His supervisor told him to dress better.

LG_54
Domain: Occupational; Topic: daily life (weekend)
Turns: 4; Words: 66
Subskill: General understanding (significant detail)

M: So how was your weekend?

W: Good. I was down in Richmond visiting my brother. We did a marathon together.

M: Really? I didn't know you ran marathons. How in the world do you find the time? With work and your family… People like you make me feel like I'm doing nothing with my life!

W: Sorry about that… But, if it's any comfort, my marathon times <u>are</u> terrible.

N: What does the man find surprising?

The woman has time to stay active.*
The woman runs faster than him.
The woman is quitting her job.
The woman visits her brother often.

LI_55
Domain: Educational; Topic: books/expense
Turns: 2; Words: 52
Subskill: implicature
Implicature type: Conversational implicature: relevance
Degree of Implicature: Medium (domain given, specific topic given)

M: One of my professors just wrote a book that I would love to read, but it costs more than a hundred dollars… How many books do they think they'll sell with prices that high?

W: Yeah, I'm not clear on the logic of it either. But I guess that's why there are libraries.

N: What does the woman suggest?

The man doesn't have to buy the book to read it.*
The man should ask his professor to lend him the book.
The book will probably not be popular.
The library won't have the book.

LI_56
Domain: Personal; Topic: objects
Turns: 3; Words: 33
Subskill: implicature
Implicature type: relevance (conversational implicature)
Degree of Implicature: high (General context: not provided; specific topic: provided; two cues)

M: I need to return this shovel to Mr. Harris. I wonder if he's home…

W: [*slightly incredulous*] Tom, that's a <u>snow</u> shovel-- we're going to the beach tomorrow!

M: [*sheepishly]* I know. I totally forgot I borrowed it.

N: What is the woman commenting on?

The man has had the shovel too long.*
The man borrows too many things.
The man has the wrong shovel.
The man forgot about their beach trip.

LI_57
Domain: Educational; Topic: academic situation (Exams)
Turns: 3; Words: 50
Subskill: implicature
Implicature type: indirect negative evaluation (conversational implicature)
Degree of Implicature: low (General context: No; specific topic yes: CUES – 3)

M: What did you think of the exam?

W: I don't know. All the short answer questions were ones that I had prepared for and made sense. But that essay question— [*emphatically*] what was that about?

M: You could give me an hour, and you know, I still don't think I could tell you.

N: How do the speakers feel about the exam?

They thought part of it was confusing.*
They weren't given enough time to complete it.
They should have studied more.
They were given the wrong questions.

LI_58
Domain: Personal; Topic: everyday life
Turns: 3; Words: 48
Subskill: implicature
Implicature type: indirect negative evaluation (conversational implicature)
Degree of Implicature: Low (General context – not given, specific context given)

W: Wow, when those airplanes fly overhead, it's pretty much impossible to hear yourself think.

M: I know. But it's really the only bad thing about this apartment. And fortunately, it's only noticeable when the windows are open.

W: Which is about half the year! I think it'd drive me crazy.

N: What has the woman commented on?

A loud noise*
A broken window
The size of the apartment
The temperature outside

LI_59
Domain: Public/Personal; Topic: parking
Turns: 4; Words: 38
Subskill: implicature
Implicature type: indirect denial (conversational implicature)
Degree of Implicature: Medium (General context – not given, specific context given)

M2: [*annoyed*] You told me it was okay to park here, Susan!

W2: Why, what's wrong?

M2: [*annoyed*] That's a parking ticket on my car.

W2: Oh. That's really unfortunate. But, if you recall, I only said I "think" it's okay to park there.

N: What does the woman claim?

She isn't to blame for the parking ticket.*
She doesn't know where the car is parked.
She didn't see the "no parking" sign.
She doesn't usually park on the street.

LI_60
Domain: personal; Topic: housing
Turns: 4; Words: 55
Subskill: implicature
Implicature type: indirect negative evaluation (conversational implicature)
Degree of Implicature: Medium (General context – not given, specific context given)

M: Did you see the sign that Mr. Daley put up near the trash cans?

W: No, what does it say?

M: Something about not putting regular garbage in the recycling bins. Though the wording is a lot less polite than that.

W: Yeah, well, I don't think politeness works for a lot of the people in this building.

N: Why did Mr. Daley put up a sign?

People are not following the rules for recycling.*
People in the building are making too much noise.
He doesn't like talking to people.
He is going to sell the building.

Appendix B: Implicature item metadata

Item Code	Domain	Number of turns	Number of words	Vocabulary frequency band	% within vocab band	Estimated strength of implicature
LI_01	Personal	2	27	5,000	96.2	Strong
LI_02	Occupational	4	52	3,000	95.9	Medium
LI_03	Educational	4	46	3,000	100	Weak
LI_04	Occupational	3	34	2,000	94.3	Strong
LI_05	Occupational	4	48	2,000	98	Weak
LI_06	Educational	3	76	3,000	94.2	Weak
LI_07	Educational	3	49	2,000	98	Strong
LI_08	Personal	5	57	2,000	93.1	Weak
LI_09	Occupational	4	54	4,000	96	Medium
LI_10	Occupational	3	34	3,000	94.7	Medium
LI_11	Personal	4	28	5,000	96.4	Strong
LI_12	Educational	2	34	3,000	97.4	Medium
LI_13	Educational	5	40	3,000	97.6	Medium
LI_14	Personal	3	48	5,000	98.1	Medium
LI_15	Occupational	4	67	3,000	98.6	Weak
LI_16	Occupational	2	39	3,000	100	Strong
LI_17	Educational	2	34	7,000	94.2	Strong
LI_18	Occupational	4	36	2,000	97.2	Weak
LI_19	Educational	3	33	1,000	97.1	Medium
LI_20	Educational	2	41	3,000	100	Strong
LI_21	Public	4	38	5,000	100	Strong
LI_22	Personal	2	21	2,000	95.2	Weak
LI_23	Educational	4	54	5,000	98.3	Medium
LI_24	Personal	2	28	3,000	100	Weak
LI_55	Educational	2	52	3,000	100	Medium
LI_56	Personal	3	33	6,000	93.11	Low
LI_57	Educational	3	50	3,000	100	Weak
LI_58	Personal	3	48	4,000	95.92	Weak
LI_59	Public	4	38	2,000	94.45	Medium
LI_60	Personal	4	55	6,000	98.11	Medium

Appendix C: Biographical questions

After giving informed consent and seeing a welcome screen, all study participants were asked to respond to the following questions (only questions 2 and 3 were mandatory).

1. Please indicate your gender.
__ Female __ Male __ No answer

2. How old are you? ____ [numeral field and mandatory]

3. What is your native language? ("Native language" is the language(s) you spoke on a daily basis within your household as a child.) ____ [open text field and mandatory]

4. At approximately what age did you first start to study or use English? ____ [numeral field]

5. For approximately how many years have you lived in an English-speaking country? (For example, Australia, Canada, or the United States) ____ [numeral field]

6. For approximately how many years have you studied English in language classes (in school or outside school)? ____ [numeral field]

7. On a scale of 1 to 5, with 1 meaning "low" and 5 meaning "high," how would you rate your overall English language ability?

__ 1 __ 2 __ 3 __ 4 __ 5 [button array]

8. On a scale of 1 to 5, with 1 meaning "low" and 5 meaning "high," how would you rate your overall English **listening** ability?

__ 1 __ 2 __ 3 __ 4 __ 5 [button array]

9. Have you ever taken the MET (the Michigan English Test)?

___Yes ____ No

If answer to question 8 was "Yes", then questions 10 and 11 were presented:

10. What score did you receive for the MET Listening section? ____ [numeral field]

11. What was your CEFR listening level on the MET? ___
__A2 __ B1 __ B2 __ C1 [button array]

Appendix D: Order and format of listening items in seven test forms

Item Code	Form 1		Form 2		Form 3		Form 4		Form 5		Form 6		Form 7	
	Position	Format	Position	Format	Position	Format	Position	Format	Position	Format	Position	Format	Position	Format
LI_01	15	CR	na	Na	38	MC	26	CR	na	na	3	MC	7	CR
LI_02	19	CR	na	Na	24	MC	22	CR	na	na	17	MC	10	CR
LI_03	16	CR	na	Na	25	MC	25	CR	na	na	16	MC	4	CR
LI_04	12	CR	na	Na	22	MC	29	CR	na	na	19	MC	3	CR
LI_05	18	CR	na	Na	32	MC	23	CR	na	na	9	MC	14	CR
LI_06	11	CR	na	Na	30	MC	30	CR	na	na	11	MC	19	CR
LI_07	10	CR	na	Na	27	MC	31	CR	na	na	14	MC	2	CR
LI_08	7	CR	na	Na	36	MC	34	CR	na	na	5	MC	11	CR
LI_09	6	CR	na	Na	35	MC	35	CR	na	na	6	MC	18	CR
LI_10	4	CR	na	Na	23	MC	37	CR	na	na	18	MC	17	CR
LI_11	40	MC	25	CR	na	na	1	MC	16	CR	na	na	na	na
LI_12	37	MC	27	CR	na	na	4	MC	14	CR	na	na	na	na
LI_13	38	MC	21	CR	na	na	3	MC	20	CR	na	na	na	na
LI_14	35	MC	24	CR	na	na	6	MC	17	CR	na	na	na	na
LI_15	34	MC	35	CR	na	na	7	MC	6	CR	na	na	na	na
LI_16	29	MC	33	CR	na	na	12	MC	8	CR	na	na	na	na
LI_17	33	MC	23	CR	na	na	8	MC	18	CR	na	na	na	na
LI_18	31	MC	30	CR	na	na	10	MC	11	CR	na	na	na	Na
LI_19	22	MC	38	CR	na	na	19	MC	3	CR	na	na	na	Na
LI_20	25	MC	39	CR	na	na	16	MC	2	CR	na	na	na	Na
LI_21	na	Na	5	MC	9	CR	na	na	36	MC	32	CR	34	MC
LI_22	na	Na	3	MC	14	CR	na	na	38	MC	27	CR	38	MC
LI_23	na	Na	2	MC	11	CR	na	na	39	MC	30	CR	40	MC
LI_24	na	Na	6	MC	6	CR	na	na	35	MC	35	CR	31	MC
LG_25	39	MC	29	CR	na	na	2	MC	12	CR	na	na	15	CR

LG_26	30	MC	32	CR	na	na	11	MC	9	CR	na	na	12	CR
LG_27	28	MC	26	CR	na	na	13	MC	15	CR	na	na	16	CR
LG_28	26	MC	28	CR	na	na	15	MC	13	CR	na	na	9	CR
LG_29	24	MC	37	CR	na	na	17	MC	4	CR	na	na	5	CR
LG_30	23	MC	36	CR	na	na	18	MC	5	CR	na	na	6	CR
LG_31	27	MC	40	CR	na	na	14	MC	1	CR	na	na	1	CR
LG_32	36	MC	34	CR	na	na	5	MC	7	CR	na	na	8	CR
LG_33	21	MC	22	CR	na	na	20	MC	19	CR	na	na	20	CR
LG_34	32	MC	31	CR	na	na	9	MC	10	CR	na	na	13	CR
LG_35	21	CR	na	Na	40	MC	21	CR	na	na	1	MC	30	MC
LG_36	13	CR	na	Na	39	MC	28	CR	na	na	2	MC	36	MC
LG_37	3	CR	na	Na	37	MC	38	CR	na	na	4	MC	23	MC
LG_38	17	CR	na	Na	34	MC	24	CR	na	na	7	MC	33	MC
LG_39	9	CR	na	Na	33	MC	32	CR	na	na	8	MC	26	MC
LG_40	8	CR	na	Na	31	MC	33	CR	na	na	10	MC	32	MC
LG_41	2	CR	na	Na	29	MC	39	CR	na	na	12	MC	29	MC
LG_42	1	CR	na	Na	28	MC	40	CR	na	na	13	MC	39	MC
LG_43	14	CR	na	Na	26	MC	27	CR	na	na	15	MC	37	MC
LG_44	5	CR	na	Na	21	MC	36	CR	na	na	20	MC	35	MC
LG_45	na	Na	18	MC	20	CR	na	na	23	MC	21	CR	na	na
LG_46	na	Na	7	MC	18	CR	na	na	34	MC	23	CR	na	na
LG_47	na	Na	1	MC	16	CR	na	na	40	MC	25	CR	na	na
LG_48	na	Na	16	MC	15	CR	na	na	25	MC	26	CR	na	na
LG_49	na	Na	19	MC	13	CR	na	na	22	MC	28	CR	na	na
LG_50	na	Na	13	MC	12	CR	na	na	28	MC	29	CR	na	na
LG_51	na	Na	14	MC	10	CR	na	na	27	MC	31	CR	na	na
LG_52	na	Na	12	MC	8	CR	na	na	29	MC	33	CR	na	na
LG_53	na	Na	4	MC	7	CR	na	na	37	MC	34	CR	na	na
LG_54	na	Na	20	MC	5	CR	na	na	21	MC	36	CR	na	na
LI_55	na	Na	17	MC	17	CR	na	na	24	MC	24	CR	21	MC

LI_56	na	Na	11	MC	19	CR	na	na	30	MC	22	CR	24	MC
LI_57	na	Na	10	MC	4	CR	na	na	31	MC	37	CR	25	MC
LI_58	na	Na	8	MC	3	CR	na	na	33	MC	38	CR	28	MC
LI_59	na	Na	9	MC	2	CR	na	na	32	MC	39	CR	27	MC
LI_60	na	Na	15	MC	1	CR	na	na	26	MC	40	CR	22	MC

Appendix E: Rasch analysis MC item results

TABLE 13.1 All MC data - uniform item codes.xlsx ZOU813WS.TXT Mar 27 2018 11:50

INPUT: 253 PERSON 60 ITEM REPORTED: 253 PERSON 60 ITEM 2 CATS WINSTEPS 3.93.2

PERSON: REAL SEP.: 1.59 REL.: .72 ... ITEM: REAL SEP.: 3.64 REL.: .93

ITEM	ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MODEL MEASURE	S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	POINT-MEASURE CORR.	POINT-MEASURE CORR. EXP	EXACT Match Obs%	Exact Match EXP%
LIM56	56	15	106	2.53	0.3	1.17	0.9	1.67	1.7	0.18	0.36	86.8	86.3
LIM03	3	11	69	2.35	0.36	1.17	0.8	1.5	1.3	0.17	0.36	81.2	85
LGM28	28	15	72	1.67	0.32	1.36	1.9	1.54	1.6	0.06	0.38	72.2	80.6
LIM12	12	17	67	1.44	0.31	1.03	0.3	1.23	0.9	0.34	0.39	76.1	77.5
LGM25	25	17	71	1.44	0.3	1.3	1.8	2.04	3.1	0.04	0.38	73.2	78.4
LGM53	53	22	73	1.38	0.29	1.19	1.4	1.28	1.2	0.28	0.44	72.6	75.3
LIM08	8	24	72	1.2	0.28	1.06	0.5	1.08	0.5	0.37	0.42	69.4	72.8
LGM39	39	34	101	1.19	0.24	0.99	0	0.94	-0.3	0.45	0.44	73.3	73.4
LGM30	30	19	64	1.17	0.3	0.94	-0.4	1.07	0.4	0.45	0.42	76.6	75.7
LGM32	32	23	71	0.96	0.28	0.77	-1.9	0.7	-1.6	0.62	0.42	84.5	74.3
LGM26	26	23	70	0.94	0.28	0.75	-2	0.68	-1.8	0.63	0.42	80	73.9
LIM24	24	40	106	0.93	0.23	0.86	-1.4	0.77	-1.6	0.58	0.46	78.3	72.5
LIM59	59	39	103	0.92	0.23	1.22	2	1.26	1.6	0.3	0.46	64.1	72.5
LIM20	20	22	67	0.91	0.29	1.11	0.8	1.14	0.8	0.32	0.42	70.1	73.8
LIM17	17	26	72	0.76	0.27	1.16	1.3	1.12	0.7	0.3	0.42	66.7	72.5
LIM23	23	43	104	0.76	0.22	0.95	-0.5	0.85	-1.1	0.52	0.46	67.3	71.2
LIM10	10	32	74	0.71	0.26	1.21	1.9	1.36	2.2	0.27	0.45	65.8	69.5
LGM46	46	31	71	0.64	0.27	0.85	-1.4	0.81	-1.2	0.57	0.46	76.1	70.2
LIM11	11	28	70	0.57	0.27	0.98	-0.1	1.21	1.3	0.41	0.43	75.7	71

LIM15	15	28	67	0.47	0.28	1.4	3.3	1.74	3.9	0.05	0.43	58.2	70.3
LIM57	57	49	105	0.46	0.22	1.09	1	1.08	0.6	0.4	0.47	66.7	70.4
LIM05	5	35	72	0.42	0.26	1.1	1	1.06	0.4	0.37	0.44	60.6	68
LIM16	16	31	72	0.4	0.26	0.8	-2.1	0.73	-2	0.61	0.43	77.8	69.6
LIM58	58	51	101	0.32	0.23	0.93	-0.8	0.87	-1	0.53	0.46	71.3	70.2
LIM09	9	37	72	0.31	0.26	1.11	1.1	1.15	1	0.36	0.45	67.6	68.7
LGM42	42	56	107	0.25	0.22	1.04	0.5	1.08	0.7	0.41	0.44	69.8	69.1
LIM55	55	50	99	0.25	0.23	1.14	1.5	1.2	1.4	0.35	0.46	66.7	69.9
LGM36	36	58	109	0.2	0.21	0.93	-0.8	0.93	-0.6	0.49	0.44	75.9	68.9
LIM22	22	50	97	0.19	0.23	1.05	0.6	1.07	0.6	0.41	0.45	67	69.7
LIM07	7	39	73	0.19	0.26	0.98	-0.2	0.92	-0.5	0.47	0.44	66.7	68.7
LIM06	6	39	73	0.18	0.26	1.14	1.3	1.09	0.7	0.34	0.44	58.3	68.6
LGM31	31	35	70	0.02	0.26	0.98	-0.1	0.99	0	0.44	0.42	67.1	67.2
LGM38	38	62	107	-0.04	0.22	0.97	-0.3	0.92	-0.6	0.46	0.43	66	69.4
LGM54	54	41	70	-0.13	0.27	0.89	-1	0.86	-0.7	0.54	0.46	78.6	71
LGM35	35	64	106	-0.18	0.22	1.02	0.3	0.97	-0.2	0.41	0.42	67.6	69.6
LGM40	40	67	110	-0.2	0.22	1	0	0.93	-0.4	0.43	0.42	66.1	69.8
LGM52	52	45	74	-0.26	0.27	0.97	-0.3	0.87	-0.7	0.49	0.45	68.9	71.5
LGM49	49	45	73	-0.31	0.27	0.92	-0.6	0.87	-0.6	0.51	0.45	75.3	71.9
LGM37	37	68	108	-0.31	0.22	0.81	-2.3	0.76	-1.7	0.56	0.41	79.4	70.2
LIM13	13	42	72	-0.35	0.26	0.99	-0.1	0.96	-0.2	0.43	0.42	63.9	67.9
LGM27	27	44	73	-0.45	0.26	0.78	-2.4	0.68	-2.2	0.62	0.41	78.1	68.4
LGM47	47	45	69	-0.46	0.28	0.96	-0.3	0.89	-0.5	0.48	0.44	76.8	72.8
LIM01	1	48	73	-0.46	0.27	0.88	-1.1	0.77	-1.2	0.51	0.4	75	71.1
LGM43	43	72	108	-0.51	0.22	1.15	1.6	1.25	1.4	0.27	0.4	67.3	71.3
LIM14	14	45	71	-0.57	0.27	1.14	1.3	1.21	1.2	0.26	0.39	64.8	69.3
LGM29	29	50	73	-0.88	0.27	0.8	-1.8	0.66	-1.8	0.58	0.39	76.7	72.9
LGM51	51	54	72	-1.07	0.3	0.81	-1.3	0.69	-1.1	0.56	0.41	81.9	78
LGM48	48	54	72	-1.08	0.3	1.09	0.6	1.29	1	0.32	0.41	76.4	77.9
LGM34	34	54	73	-1.2	0.29	0.8	-1.5	0.72	-1.2	0.55	0.37	82.2	76.7

LIM60	60	81	104	-1.22	0.26	0.87	-1	0.64	-1.4	0.51	0.38	81.7	79.4
LGM41	41	86	108	-1.3	0.26	0.96	-0.2	0.98	0	0.36	0.33	80.4	80
LGM50	50	56	71	-1.34	0.32	0.96	-0.1	1.05	0.3	0.4	0.39	83.1	80.6
LIM04	4	57	71	-1.35	0.32	0.77	-1.4	0.54	-1.5	0.54	0.34	82.9	80.7
LGM44	44	85	105	-1.39	0.27	0.92	-0.5	1.21	0.8	0.36	0.32	82.7	81.2
LIM18	18	55	70	-1.47	0.31	0.84	-1	0.65	-1.2	0.52	0.36	81.4	80.2
LIM02	2	60	73	-1.49	0.33	0.98	0	0.79	-0.5	0.36	0.32	81.9	82.4
LIM21	21	91	108	-1.74	0.28	0.91	-0.5	0.62	-1.1	0.44	0.33	84.3	84.7
LGM33	33	62	72	-2.1	0.36	0.95	-0.1	1.32	0.8	0.3	0.31	87.5	86.4
LIM19	19	63	73	-2.11	0.36	0.96	-0.1	0.79	-0.4	0.35	0.31	87.7	86.6
LGM45	45	65	74	-2.13	0.38	1	0.1	1.24	0.6	0.29	0.32	87.8	87.8
MEAN		45	8.25	0	0.27	1	0	1.02	0			74.2	74.3
P.SD		18.7	16.2	1.07	0.04	0.15	1.2	0.3	1.3			7.6	5.6

Appendix F: Rasch analysis CR item results

TABLE 13.1 All CR data - uniform item codes.xlsx ZOU568WS.TXT Mar 27 2018 11:43

INPUT: 254 PERSON 60 ITEM

REPORTED: 254 PERSON 60 ITEM 3 CATS WINSTEPS 3.93.2

PERSON: REAL SEP.: 2.07 REL.: .81

ITEM: REAL SEP.: 4.03 REL.: .94

ITEM STATISTICS: MEASURE ORDER

ITEM	ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MODEL MEASURE	S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	POINT-MEASURE CORR.	POINT-MEASURE CORR. EXP	EXACT Match Obs%	Exact Match EXP%
LIC56	56	5	68	2.69	0.41	1.33	0.7	1.42	0.7	0.2	0.23	95.6	93.9
LIC22	22	17	62	1.56	0.24	1.21	0.9	1.42	1	0.3	0.4	69.4	74.5
LGC31	31	40	99	1.11	0.16	0.98	-0.1	0.92	-0.2	0.44	0.46	68.7	66.7
LGC39	39	26	52	0.98	0.22	0.88	-0.5	0.7	-0.7	0.6	0.53	71.2	65.3
LIC59	59	34	71	0.92	0.18	0.95	-0.2	0.78	-0.7	0.56	0.49	62	64.3
LIC08	8	53	90	0.88	0.16	1.03	0.3	0.83	-0.6	0.59	0.54	58.9	57.9
LIC24	24	38	72	0.81	0.18	1.08	0.5	0.91	-0.3	0.46	0.5	54.2	59.1
LIC58	58	34	58	0.73	0.19	0.86	-0.8	0.71	-1.1	0.67	0.51	55.2	54
LIC20	20	34	59	0.72	0.19	1.18	1.1	1.21	0.9	0.33	0.48	52.5	55.5
LIC60	60	40	61	0.65	0.18	0.93	-0.4	0.9	-0.3	0.56	0.51	50.8	49.9
LIC16	16	39	56	0.61	0.18	0.84	-1	0.69	-1.4	0.69	0.5	50	48.7
LIC09	9	73	102	0.53	0.14	1.18	1.4	1.3	1.4	0.53	0.55	40.2	48.1
LIC13	13	48	69	0.53	0.16	0.72	-2.2	0.62	-2.1	0.7	0.5	56.5	49.1

LIC07	7	72	97	0.51	0.14	0.96	-0.3	0.89	-0.4	0.64	0.55	45.4	47.3
LIC03	3	65	87	0.5	0.15	1.15	1.1	0.99	0	0.54	0.55	46	47
LIC11	11	49	68	0.48	0.16	1.07	0.6	0.98	0	0.52	0.51	47.1	47.1
LIC05	5	77	101	0.46	0.14	1.45	3.4	1.93	3.8	0.26	0.54	41.6	45.5
LIC15	15	45	61	0.44	0.17	0.58	-3.3	0.62	-1.9	0.64	0.53	65.6	48.4
LIC17	17	50	67	0.37	0.16	1.39	2.6	1.49	2.4	0.25	0.49	34.3	43.7
LIC57	57	56	73	0.32	0.16	0.71	-2.3	0.82	-0.9	0.34	0.53	52.1	47.4
LGC53	53	57	71	0.28	0.16	0.77	-1.8	0.72	-1.6	0.65	0.52	53.5	48.2
LGC25	25	88	104	0.23	0.13	0.74	-2.6	0.69	-2.3	0.64	0.53	50	44.4
LGC26	26	90	100	0.19	0.13	0.92	-0.7	0.84	-1.1	0.53	0.52	51	44.4
LGC41	41	61	68	0.19	0.16	1.06	0.5	0.9	-0.4	0.59	0.55	39.7	45.9
LIC12	12	64	74	0.18	0.15	1.12	0.9	1.14	0.9	0.38	0.5	41.9	41
LGC32	32	94	108	0.17	0.13	0.83	-1.7	0.96	-0.2	0.45	0.53	46.3	44.4
LGC54	54	60	69	0.16	0.16	0.83	-1.3	0.71	-1.7	0.71	0.53	47.8	46
LIC04	4	91	94	0.08	0.14	1.19	1.6	1.19	1	0.54	0.57	41.5	49
LIC23	23	63	62	-0.02	0.17	0.82	-1.4	0.89	-0.5	0.57	0.54	50	44.2
LIC06	6	106	104	-0.04	0.13	0.79	-2.1	0.72	-1.7	0.64	0.55	51	47.2
LIC10	10	93	93	-0.04	0.14	1.19	1.6	1.11	0.6	0.52	0.56	43	47.9
LGC30	30	90	83	-0.04	0.14	0.95	-0.4	0.84	-1	0.63	0.51	43.4	41.8
LGC38	38	73	70	-0.1	0.16	0.98	-0.1	0.81	-0.8	0.62	0.56	47.1	47.9
LGC36	36	72	68	-0.14	0.16	0.97	-0.2	0.89	-0.4	0.58	0.56	47.1	47.8
LIC14	14	68	63	-0.14	0.16	1.43	2.9	1.67	3.1	0.24	0.5	33.3	41.4
LGC52	52	76	74	-0.15	0.15	0.86	-1.1	0.89	-0.6	0.52	0.51	44.6	43.1
LIC21	21	77	72	-0.19	0.15	1.12	1	1.23	1.3	0.4	0.52	43.1	43.9
LIC18	18	82	72	-0.24	0.15	0.77	-2.1	0.87	-0.7	0.32	0.48	55.6	39.9
LGC47	47	82	70	-0.32	0.16	0.8	-1.6	0.87	-0.7	0.57	0.5	48.6	42.1
LGG37	37	83	70	-0.36	0.16	1.14	1	1.66	2.3	0.44	0.56	50	48.8
LGC42	42	86	71	-0.39	0.16	1.05	0.4	0.94	-0.2	0.51	0.55	43.7	48.2
LGC51	51	81	67	-0.4	0.16	0.81	-1.4	0.82	-0.9	0.6	0.51	56.7	43.8
LGC28	28	116	90	-0.41	0.14	1.04	0.4	0.97	-0.1	0.56	0.49	46.7	46.1

LGC33	33	133	109	-0.42	0.13	0.94	-0.6	0.94	-0.4	0.43	0.5	48.6	45.1
LGC29	29	130	100	-0.52	0.13	1.04	0.4	1.21	1.2	0.43	0.47	44	44.7
LGC46	46	88	69	-0.53	0.16	1.34	2.4	1.58	2.4	0.27	0.49	36.2	43.7
LGC35	35	85	65	-0.57	0.18	1.12	0.8	1.11	0.5	0.49	0.54	46.2	50.2
LIC02	2	135	102	-0.63	0.14	0.9	-0.7	0.76	-1.1	0.64	0.54	51	51
LIC19	19	99	73	-0.66	0.16	0.95	-0.3	0.77	-1	0.6	0.46	49.3	47.8
LIC01	1	143	106	-0.68	0.14	1.26	1.9	1.64	2.5	0.36	0.53	44.3	51
LIC55	55	97	72	-0.68	0.16	1.68	4.1	2.13	3.9	0.13	0.49	37.5	48
LGC27	27	145	104	-0.71	0.14	0.71	-2.7	0.63	-2.1	0.66	0.46	50	48.5
LGC40	40	96	70	-0.73	0.18	1.01	0.1	0.79	-0.7	0.58	0.54	60	53.3
LGC45	45	106	75	-0.83	0.16	1.12	0.9	1.02	0.2	0.48	0.47	48	52
LGC50	50	106	73	-0.91	0.17	0.88	-0.8	0.73	-1	0.53	0.46	47.9	54.1
LGC48	48	103	70	-0.93	0.18	0.95	-0.3	0.75	-0.9	0.6	0.47	51.4	56.4
LGC49	49	110	75	-0.94	0.17	0.72	-2	0.62	-1.6	0.6	0.45	65.3	54.8
LGC44	44	114	74	-1.15	0.19	0.98	0	1.23	0.7	0.51	0.51	70.3	67.4
LGC34	34	186	107	-1.56	0.18	1.09	0.5	0.83	-0.4	0.4	0.34	79.4	76.1
LGC43	43	130	74	-1.84	0.24	1.24	1	1.43	0.9	0.3	0.42	79.7	78.3
MEAN		79.2	78.5	0	0.17	1.01	0	1.01	-0.1			51.7	51.2
P.SD		34.5	15.8	0.75	0.04	0.21	1.5	0.34	1.4			11.5	10.1

Appendix G: General item metadata

Item Code	Domain	Number of turns	Number of words	Vocabulary frequency band	% within vocab band
LG_25	Personal	5	70	6,000	96.1
LG_26	Occupational	4	65	2,000	97
LG_27	Educational	2	46	3,000	98
LG_28	Personal	4	45	5,000	95.8
LG_29	Personal	3	46	3,000	96
LG_30	Educational	3	50	3,000	98
LG_31	Educational	5	51	3,000	100
LG_32	Educational	3	48	2,000	90
LG_33	Personal	4	43	1,000	100
LG_34	Public	3	53	2,000	94.12
LG_35	Occupational	4	64	4,000	96.97
LG_36	Personal	4	62	2,000	98.25
LG_37	Educational	4	68	5,000	95.78
LG_38	Educational	3	63	5,000	96.68
LG_39	Occupational	4	52	4,000	96.15
LG_40	Occupational	4	60	5,000	100
LG_41	Educational	4	48	4,000	100
LG_42	Personal	3	39	6,000	94.12
LG_43	Educational	3	53	2,000	92.85
LG_44	Public	4	64	5,000	100
LG_45	Public	4	61	3,000	96.67
LG_46	Occupational	4	53	3,000	96
LG_47	Educational	3	55	4,000	96.07
LG_48	Occupational	4	60	1,000	98.28
LG_49	Occupational	4	55	4,000	96
LG_50	Personal	4	56	3,000	100
LG_51	Public	4	64	3,000	95.39
LG_52	Occupational	3	64	3,000	98.36
LG_53	Occupational	3	42	2,000	95.23
LG_54	Occupational	4	66	5,000	96.93

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*(1), 30–60.

Alderson, J. C. (Ed.) (2002). *Case studies in the use of the Common European Framework*. Strausbourg: Council of Europe.

Alderson, J. C., (2005). *Diagnosing Foreign Language Proficiency.* London, UK: Continuum.

Alderson, J. C., Figueras, N., Kuijer, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the common European framework of reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly, 3*(1), 3–30.

American Education Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing.*

American Council on the Teaching of Foreign Languages (2012). *ACTFL Proficiency Guidelines 2012.* Alexandria, VA: ACTFL.

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning 62*(2): 49–78.

Bae, J. & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English Two-Way Immersion Program. *Language Testing, 15*(3) 380–414.

Bachman, L. F., (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. & Palmer A. S., (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Baddeley, A. D., (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences,* 4(11), 417–423.

Baddeley, A. D., (2007). *Working Memory, Thought, and Action.* Oxford: Oxford University Press.

Baddeley, A. D., (2010). Working memory. *Current Biology*, *20*(4), R136–R140.

Baddeley, A. D., & Hitch, G., (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47–90). New York: Academic Press.

Blakemore, D. (1992). *Understanding Utterances: An Introduction to Pragmatics.* Oxford, UK: Blackwell Publishers.

Bloomfield, A. N., Ross, S. J., Masters, M. C., Gynther, K. & O'Connell, S. P. (2014). How Does Foreign Language Proficiency Change Over Time? Results of Data-Mining Official Test Records. In J. Connor-Linton & L. Amoroso (Eds.) *Measured Language: Quantitative Studies of Acquisition, Assessment, and Variation* (pp. 199–212). Washington, DC: Georgetown University Press.

Bloomfield, A. N., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). What makes listening difficult? Retrieved from https://apps.dtic.mil/docs/citations/ADA550176 on March 17, 2018. University of Maryland Center for Advanced Study of Language.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review 111(4)*, 1061-1071.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. Lissitz (Ed.) *The Concept of Validity* (pp. 135–170). Charlotte NC: Information Age Publishing, Inc.

Bodie, G. D., Worthington, D., Imhof, M., & Cooper, L. O. (2008). What would a unified field of listening look like? A proposal linking past perspectives and future endeavors. *International Journal of Listening, 22,* 103–122.

Bond, T. G. & Fox, C. M. (2001). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bouton, L. (1988). A cross-cultural study of ability to interpret implicatures in English. *World Englishes, 17*: 183–196).

Bouton, L. F. (1994a). Conversational implicature in the second language: learned slowly when not deliberately taught. *Journal of Pragmatics, 22*: 157–167.

Bouton, L. F. (1994b). Can NNS skill in interpreting implicature in American English be improved through explicit instruction? A pilot study. *Pragmatics and Language Learning, 5*: 88–109.

Bouton, L. F. (1999). Developing non-native speaker skills in interpreting conversational implicatures in English: Explicit teaching can ease the process. In E. Hinkel (ed.), *Culture in Second Langauge Teaching and Learning* (pp. 47–70). Cambridge, UK: Cambridge University Press.

Buck, G., (1991). The testing of listening comprehension: an introspective study. *Language Testing, 8*(1): 67–91.

Buck, G., (2001). *Assessing Listening*. Cambridge, UK: Cambridge University Press.

Brehmer, Y., Li, S. C., Muller, V., von Oertzen, T., & Linderberger, U. (2007). Memory plasticity across the life span: Uncovering children's latent potential. *Developmental Psychology, 43,* 465–478.

Brown, T. A. (2006) *Confirmatory Factor Analysis for Applied Research.* New York, NY: The Guildford Press.

Brunfaut, T. & Révész, A. (2011). *EAP listening task difficulty: The impact of task variables, working memory and listening anxiety.* Paper presented at the Language Testing Research Colloquium. Ann Arbor, MI, June 23–25.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod maxtrix. *Psychological Bulletin, 56*(2): 81–105.

Campbell, S. G., Hughes, M. M., Smith B. K., Meyers, J. H., O'Connell, S. (2012). Reliability and validity of the English listening comprehension test: Creation and evaluation of a measure of first-language listening comprehension. University of Maryland Center for Advanced Study of Language, Technical Report 82114 8.1

Cambridge Michigan Language Assessments (2012). Setting Cut Scores on the Common European Framework of Reference for the Michigan English Test. Technical Report. Retrieved from: http://www.cambridgemichigan.org/wp-content/uploads/2014/12/MET_StandardSetting.pdf on 30 August 2015.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller (Ed.) *Issues in Language Testing Research* (pp. 333–342) Rowley, MA: Newbury House.

Canale, M. and Swain, M. (1981). A theoretical framework for communicative competence. In A. S. Palmer, P. J. M. Groot, and G. A. Trosper (Eds.) *The Construct Validation of Tests of Communicative Competence* (pp. 31–36), Washington, DC: Teachers of English to Speakers of Other Languages.

Caplan, D., Waters, G., & Dede, G., (2007). Specialized Verbal Working Memory for Language Comprehension. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, J. N. Towse (Eds) *Variation in working memory* (pp. 272–302), Oxford: Oxford University Press.

Carr, N. & Kunnan, A. (2016). Feasibility of adapting a human-scored short-answer reading test to computer-automated scoring. Paper presented at 2016 American Association of Applied Linguistics, Orlando, Florida.

Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple… *Language Testing,* 29(1), p. 19-27.

Cheung, H., (1996). Nonword span as a unique predictor of second-language vocabulary learning. *Developmental Psychology*, *32,* 867–873.

Clark, M. (2007). Listening Placement Test Development and Analysis from a Rasch Perspective (unpublished doctoral dissertation). University of Hawai'i-Manoa, Hawai'i, USA.

Clark, M., Wayland, S., Osthus, P., Brown, K. G., & Castle, S. (2014). The effects of note taking on foreign language listening comprehension. Retrieved from http://www.govtilr.org/publications/notetaking.pdf University of Maryland Center for Advanced Study of Language, Technical Report.

Conway, A. R. A., Kane, M. J., Bunting, M., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769-786.

Conway, A. R. A, Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. N., eds (2007). *Variation in Working Memory.* Oxford: Oxford University Press.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment – A Manual*. Strausbourg: Council of Europe.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Companion Volume with New Descriptors.* Strausbourg: Council of Europe.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin, 104,* 163–191.

Cowan, N., (1999). An embedded processes model of working memory In A. Miyake & P. Shah (eds.) *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control,* (pp. 62–101). New York, NY: Cambridge University Press.

Cowan, N. (2005). Working Memory Capacity: Essays in Cognitive Psychology. New York: Psychology Press.

Cronbach L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Crossley, S. (2016). Assessing lexical proficiency using the English Lexicon Project and a bit more. Paper presented at 2016 American Association of Applied Linguistics, Orlando, Florida.

Cruse, A. (2000). *Meaning in Language: An Introduction to Semantics and Pragmatics.* Oxford, UK: Oxford University Press.

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450–466.

Daneman, M. & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 561–583.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422–433.

Dunkel, P., Henning, G., Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal, 77*(2), 180–191.

Edwards, R. (2011). Listening and message interpretation. *International Journal of Listening, 25,* 47-65.

Ellis, N. C., (2001). Memory for Language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33–68). New York: Cambridge University Press.

Engle, R. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science 11*(1), 19-23.

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R., Mollaun, P., Nissan, S., Powers, D. E., & Schedl, M. (2008) In C. Chapelle, M. K. Enright, & J. M. Jamieson (Eds), *Building a Validity Argument for the Test of English as a Foreign Language*. New York: Routledge.

Ericsson, K. A. and Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211–245.

Eom, M., (2008). Underlying factors of MELAB listening constructs. *Spaan Fellow Working Papers in Second or Foreign Language Assessment,* Volume 6, 77–94.

Field, J. (2008). *Listening in the Language Classroom.* Cambridge, UK: Cambridge University Press.

Field, J. (2013). Cognitive Validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining Listening: Research and Practice in Assessing Second Language Listening* (pp. 77–151). Cambridge, UK: Cambridge University Press.

Freedle, R. & Kostin, I. (1996). The prediction of TOEFL listening comprehension item difficulty for mini-talk passages: implications for construct validity. (TOEFL Research Report No. 96-29). Princeton, NJ: Educational Testing Service.

Freedle, R. & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, 16(1) 2–32.

French, L. M. *&* O'Brien, I. *(*2008*).* Phonological memory and children's second language grammar learning*. Applied Psycholinguistics, **29**,* 463–*4*87.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253–266.

Gathercole, S. E., (ed) (1996). *Models of Short-term Memory.* Hove, UK: Erlbaum.

Gathercole, S. E. & Pickering, S. J., (1999). Estimating the capacity of phonological short-term memory. *International Journal of Psychology*, 34, pp 378–382.

Gathercole, S. E., Service, E., Hitch, G., Adams, A. M., & Martin, A. J., (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology*, 13, pp 65–77.

Genesee F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English Language Learners in U.S. Schools: An Overview of Research Findings, *Journal of Education for Students Placed at Risk (JESPAR)*, 10:4, 363–385.

Grice, P. (1989). *Studies in the Ways of Words.* Cambridge, US: Harvard University Press.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences.* Boston, MA: Houghton Mifflin Company.

Harrington, M. & Sawyer, M., (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition, 14*(1), 25–38.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Hummel, K. M. (2009). Aptitude, phonological memory, and second language proficiency in non-novice adult learners. *Applied Psycholinguistics 30*(2), 225–249.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride and J. Holmse (Eds.), *Sociolinguistics* (269–293) Harmondsworth, UK: Penguin.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*, p. 229-249.

In'nami, Y. and Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing,* 26(2) 219-244.

Juffs, A. & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching 44*(2), 137-166.

Kaftandjieva, F. & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study, in Alderson, J. C. (Ed) *Case Studies in the Use of the Common European Framework,* Strasbourg: Council of Europe, pp. 106–129.

Kane, M. (1992). An argument based approach to validity. *Psychological Bulletin, 112(3),* p. 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational Measurement*, 3rd ed. (pp. 13-103). New York, NY: Macmillan.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. Lissitz (Ed.) *The Concept of Validity* (39-64). Charlotte, NC: Information Age Publishing, Inc.

Kane, M. (2011). Validating score interpretations and uses: Messick lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing,* 29(1), 3-17.

Kasper, G. (1984) Pragmatic comprehension in learner-native speaker discourse. *Language Learning, 34*(4), 1-20.

Kasper, G. and Rose, K. R. (2002). *Pragmatic Development in a Second Language.* Malden, MA: Blackwell Publishing.

Kenny, D. A. (1979). *Correlation and causality.* New York, NY: Wiley-Interscience.

Kenny, D. A. (2012, March). Multitrait multimethods matrix: Definitions and introduction. Retrieved from http://davidakenny.net/cm/mtmm.htm on December 1, 2018.

Kenny, D. A. & Kashy, D. A. Analysis of multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Kintsch, W. and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review,* 85(5), p. 363-394.

Kramsch, D. (2009). *The Multilingual Subject.* Oxford, UK: Oxford University Press.

Kostin, I. (2004). Exploring item characteristics that are related to the difficulty of TOEFL dialogue items. (TOEFL Research Report No. 79). Princeton, NJ: Educational Testing Services.

Kormos, J. & Sáfár, A (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition 11*(2), 261–271.

Kyllonen, P. C. and Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity! *Intelligence* 14, 389-433.

Linacre, J. M. (1994). Sample Size and Item Calibration Stability. *Rasch Measurement Transactions*, Volume 7.

Linacre, J. M. (2017). A User's Guide to WINSTEPS®: MINISTEP Rasch-Model Computer Programs 4.0.0

Linacre, J. M. (2019). A User's Guide to WINSTEPS®: MINISTEP Rasch-Model Computer Programs 4.4.2

Linacre, J. M. (2018, September). *Detecting multidimensionality in Rasch data using Winsteps Table 23.* Retrieved from https://www.youtube.com/watch?v=sna19QemE50&feature=youtu.be

Linck, J. A., Osthus, P., Koethe, J. T., & Bunting, M. F. (2014). Working memory and second language production and comprehension: a meta-analysis. *Psychonomic Bulletin Review 21*(4), 861-883.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning, 63,* 530-566.

Lissitz, R. W. and Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437-448.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing 24*(4) 489–515.

Lomax, R. G., (2007). *Statistical Concepts: A Second Course.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010. Exploring the Relationship Between Modified Output and Working Memory Capacity. *Language Learning, 60(3),* 501–533.

McDonald, J. L., (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language, 55*(3), 381–401.

McNamara, T. (1996). *Measuring Second Language Performance.* Harlow, UK: Longman.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher,* 18(2), p. 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* 23(12), 13-23.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50(9), p. 741-749.

Michigan Language Assessments (2019). *MET.* Retrieved from https://michiganassessment.org/test-takers/tests/met/ on May 19, 2019.

Michigan Language Assessments (2014, July). *ECPE Speaking Rating Scale.* Retrieved from https://michiganassessment.org/wp-content/uploads/2014/11/ECPE-Rating-Scale-Speaking.pdf on November 17, 2018.

Miyake, A. & Friedman, N. P., (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne (Eds.), *Foreign Language Learning* (pp. 339–364). London: Lawrence Erlbaum Associates.

Nielson, K. B. (2014). Can planning time compensate for individual differences in working memory capacity? *Language Teaching Research*, 18(3), 272–293.

Nissan, S., DeVincenzi, F., and Tang L. (1996). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. (TOEFL Research Report 51). Princeton, NJ: Educational Testing Service.

North, B. (2000). *The Development of a Common European Framework Scale of Language Proficiency,* New York: Peter Lang.

North, B. (2014). *The CEFR in Practice.* Cambridge, UK: Cambridge University Press.

O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B., (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics,* 27, 377–402.

Papageorgiou, S. (2009). *Setting Performance Standards in Europe: The Judges' Contribution to Relating Language Examinations to the Common European Framework of Reference.* Frankfurt: Peter Lang

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research.* Thousand Oaks, CA: Sage Publications, Inc.

Purpura, J. (2004). *Assessing Grammar*. Cambridge, UK: Cambridge University Press.

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing, 23(2)* 229-256.

Roever, C. (2013). Testing implicature under operational conditions. In S. Ross & G. Kasper (Eds). *Assessing Second Language Pragmatics.* Basingstoke, UK: Palgrave MacMillan.

Rost, M. (2011). *Teaching and Researching Listening* (second edition). Harlow, UK: Pearson Education Limited.

Rupp, A. A., Garcia, P. & Jamieson, J., (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing, 1*(3 & 4), 185–216.

Sawaki Y. & Nissan, S. (2009). Criterion-related validity of the TOEFL iBT listening section. (TOEFL Research Report No. 09-02). Princeton, NJ: Educational Testing Service.

Sawyer, M. & Ranta, L. (2001). Aptitude, individual differences, and instructional design (pp. 319–353). In P. Robinson (ed) *Cognition and Second Language Instruction.* Cambridge: Cambridge University Press.

Song, M. Y., (2008). Do Divisible Sub-skills Exist in Second Language (L2) Comprehension? A structural equation modeling approach. *Language Testing, 25*(4), 435–464.

Sperber, D. & Wilson, D. (1995). *Relevance: Communication and Cognition* (2nd edition). Oxford, UK: Blackwell.

Taguchi, N. (2005). Comprehending implied meaning in English as a foreign language. *The Modern Language Journal,* 89, 543–562.

Taguchi, N. (2007). Development of speed and accuracy in pragmatic comprehension in English as a foreign language. *TESOL Quarterly,* 41(2): 313-338.

Taguchi, N. (2008). The effect of working memory, semantic access, and listening abilities on the comprehension of conversational implicatures in L2 English. *Pragmatics & Cognition,* 16(3) 517-539

Taguchi, N. (2009). Corpus-informed assessment of comprehension of conversational implicatures in L2 English. *TESOL Quarterly,* 43(4): 738-749.

Taguchi, N. & Roever, C. (2017). *Second Language Pragmatics.* Oxford, UK: Oxford University Press.

Vafaee, P. (2016). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening comprehension (unpublished doctoral dissertation). University of Maryland, College Park, MD, USA.

Wayland, S., Saner, L., O'Connell, S., Linck, J., Kramasz, D., Gynther, K., Bloomfield, A., & Ralph, A. (2013). The long and the short of it: Passage length and information density insecond language listening comprehension. https://www.casl.umd.edu/wp-content/uploads/2016/02/THE-LONG-AND-THE-SHORT-OF-IT-2013.pdf University of Maryland Center for Advanced Study of Language, Technical Report.

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*. Volume 2. English Language Institute, University of Michigan.

Weir, C. J. (2005a). Limitations of the common European framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.

Weir, C. (2005b). *Language Testing and Validation.* Basingstoke, UK: Palgrave Macmillan.

Williams, J. N. (2012). Working memory and SLA (pp. 427–441). In S. Gass and A. Mackey (Eds.). *Handbook of Second Language Acquisition.* Routledge.

Wilson, D. and Sperber, D. (2012). *Meaning and Relevance.* Cambridge, UK: Cambridge University Press.

Wright, B. D. and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370–371.

Wu, Y. (1998). What do tests of listening comprehension test? A retrospection study of EFL test takers performing a multiple-choice task. *Language Testing, 15*(1): 21–44.