AGE, TASK CHARACTERISTICS, AND ACOUSTIC INDICATORS OF ENGAGEMENT: INVESTIGATIONS INTO THE VALIDITY OF A TECHNOLOGY-ENHANCED SPEAKING TEST FOR YOUNG LANGUAGE LEARNERS

by

Edward Paul Getman

Dissertation Committee:

Professor James Enos Purpura, Sponsor
Professor Payman Vafaee

Approved by the Committee on the Degree of Doctor of Education

Date February 12, 2020

Submitted in partial fulfillment of the
Requirements for the Degree of Doctor of Education in
Teachers College, Columbia University

2020

ABSTRACT

AGE, TASK CHARACTERISTICS, AND ACOUSTIC INDICATORS OF ENGAGEMENT: INVESTIGATIONS INTO THE VALIDITY OF A TECHNOLOGY-ENHANCED SPEAKING TEST FOR YOUNG LANGUAGE LEARNERS

Edward Paul Getman

Despite calls for engaging assessments targeting young language learners (YLLs) between 8 and 13 years old, what makes assessment tasks engaging and how such task characteristics affect measurement quality have not been well studied empirically. Furthermore, there has been a dearth of validity research about technology-enhanced speaking tests for YLLs. Thus, the purpose of the current study was to explore relationships among examinee age, task characteristics, engagement, and performance in the context of gathering evidence to back a test validity argument. Following a mixed-methods approach, the investigations involved over 400 YLLs in 11 countries who responded to *TOEFL Primary*® Speaking test tasks (Educational Testing Service, 2013). Results from many-facet Rasch measurement revealed that, in terms of evaluation claims, tasks and raters functioned well across examinee age groups. Generalizability theory was then applied to confirm that examinees accounted for most of the score variance and that current test form configurations maximize score dependability. Fischer's (1973, 1995) linear logistic test model results helped explain that vocabulary support, novelty, and video animation increased task difficulty, while topical choice did not. Lastly, acoustic measures of harmonicity and shimmer from the spoken responses served as indicators of engagement in a structural model showing that topical choice and novelty promoted engagement; these findings were triangulated by retrospective verbal

reports from eight YLLs. Results point to the importance of including engagement in theoretical models of language performance. Also, a taxonomy of task characteristics that may support engagement is proposed to help drive a research agenda and inform test development.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support of Educational Testing Service, my employer for over ten years. I also wish to thank my advisor, Dr. James Purpura, for his guidance and insights throughout my doctoral studies at Teachers College, Columbia University. In addition, I am grateful to the other members of my dissertation committee, Dr. Payman Vafaee, Dr. Vivian Lindhardsen, and Dr. Ioana Literat, for their thoughtful and thought-provoking feedback on my work. I also owe a debt of gratitude to many colleagues in the field for their unwavering encouragement. Special thanks goes to the participants in my study and to the people who contributed to data collection efforts, especially Amy Cellini, Mitch Ginsburgh, Shawn Steinhart, Hee Jin Kim, Seuk Young Jang, and Dr. Jeong Hyun Kim. I am also appreciative of Dr. James Purpura and Christine Liddie for reviewing earlier drafts of this work. Of course, any errors or omissions are my own.

E.P.G.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

The population of young language learners (YLLs), defined herein as learners of a second or foreign language between the ages of approximately 8 and 13 years, is growing rapidly around the world (Butler, 2016). In the United States, more than one in five school-aged children speak a language other than English at home (Ryan, 2013), and that number is climbing. Over a six-year period, the number of public-school students who were identified as English language learners (ELLs) grew by over a million (National Center for Education Statistics, 2004), and several states had experienced at least a 300% boom in the ELL population in just ten years (Payán & Nettles, 2008). Similar findings have been observed around the world. In fact, foreign language instruction in primary schools has become part of education policy throughout much of Europe and elsewhere (Kubanek-German, 1998). In 22 countries—over a third of the countries surveyed—English instruction begins early in primary school (Rixon, 2013). In Europe, for example, the percentage of primary school students learning English has risen over 10% in just five years (Baïdak, Borodankova, Kocanova, & Motiejunaite, 2012). In Japan, the percentage of children learning English increased 15% in the same amount of time (Graddol, 2006). There is no sign that the growth in the population of YLLs will slow down anytime soon.

The expansion in language learning among children around the world is mirrored by a growing interest in developing language assessments for them (McKay, 2006; Rea-Dickins, 2000; Taylor & Saville, 2002; Zangl, 2000). In particular, there is a need for high-quality speaking assessments given the central role that oral language tends to play in the personal and academic lives of YLLs (Cameron, 2001, 2003; McKay, 2006). Much of the literature on speaking assessments is focused on classroom-based formative assessments (e.g., Davison & Leung, 2009;

Leung & Mohan, 2004; Teasdale & Leung, 2000). However, teachers and schools believe that formal, external assessments are still important (Blanco & Howden, 2011; Cameron, 2003). Although continually gathering information about students' oral language performances to inform classroom instruction may be sound pedagogically, there can be much variability in assessment practice from teacher to teacher and context to context (Butler, 2009; Cheng, Rogers, & Hu, 2004). Large-scale assessments often introduce better uniformity of measurement across different contexts (Kunnan, 2008).

Several prominent large-scale tests have been developed to specifically assess the oral language proficiency of YLLs. These tests include the *Cambridge English: Young Learners* tests (Cambridge ESOL, 2007), the *Spoken ESOL* for young learners examinations (City & Guilds, 2008), the *PTE Young Learners* tests (Pearson, 2009), and the *TOEFL Primary*® Speaking test (Educational Testing Service, 2013). All of these internationally recognized tests are administered and scored by trained interviewers except the *TOEFL Primary* Speaking test, which is computer- or tablet-delivered with responses being recorded for later human scoring. In the United States, computer-delivered assessments include the *ACCESS for ELLs 2.0* ("WIDA," n.d.) and the *English Language Proficiency Assessments for the 21st Century* ("ELPA21," n.d.) that measure YLL speaking proficiency in many different states.

Despite the growing prevalence of such large-scale assessments, most published studies about the validity of speaking tests involve adult or adolescent language learners, not YLLs. However, YLLs present with a variety of specific traits and needs that set them apart from adolescent and adult language learners (Bailey, 2008; Inbar-Lourie & Shohamy, 2009; McKay, 2006). For example, YLLs are marked by differences in the processes of second language acquisition, by still-developing cognitive capacities, by affective factors, and by experiential limitations. Therefore, results from studies of adult language learners do not necessarily apply to children (Oliver, 1998). The dearth of research on large-scale speaking tests for YLLs is

concerning given the ever-expanding need to measure their learning progress, to cater instructional programming for them, and to make policy decisions involving them. The goal of the current study is to help fill these gaps in the literature.

Given the paucity of empirical research into how to best measure YLLs' speaking and other language abilities, experts (e.g., Bailey, 2008; Hasselgreen, 2005; Taylor & Saville, 2002; Wolf & Butler, 2017) have proposed several guidelines for the development of assessments for them. These guidelines recommend that assessments be beneficial for children and provide meaningful information to support further learning (Hauck, Wolf, & Mislevy, 2013; McKay, 2006; Shepard, 1994; Taylor & Saville, 2002). Assessments should highlight the capabilities of YLLs instead of their deficits (Hasselgreen, 2005; Muñoz, 2012; Taylor & Saville, 2002), and the development and experiences of young learners should be taken into consideration (Bailey, 2008; Kubanek-German, 1998; Muñoz, 2012; Shepard, 1994; Taylor & Saville, 2002; Wolf & Butler, 2017). In addition to constituting good learning activities in themselves (Hasselgreen, 2005; Muñoz, 2012), assessments should also make use of a wide variety of tasks and measures (Espinosa, 2012; Hasselgreen, 2005; Pitoniak et al., 2009; Shaaban, 2001; Traphagan, 1997; Zangl, 2000).

However, perhaps the most common guideline is that test tasks should be engaging (e.g., Hasselgreen, 2000, 2005; Hauck et al., 2013; McKay, 2006; Taylor & Saville, 2002; Wolf & Butler, 2017). At first glance this dictum may seem self-evident, but a closer inspection reveals a chain of assumptions that have not been well investigated empirically. For example, first is an assumption that test developers intuitively know what makes tasks engaging for YLLs. However, no one has issued a credible taxonomy of task characteristics that may support engagement, let alone confirmed that these characteristics really do support engagement. Secondly, although many people would probably say they can recognize engagement when they see it, objective and dependable indicators of engagement have remained elusive. Thirdly, the relationship between

task engagement, or at least the task characteristics that support it, and task performance has not been clear. These problems provided the motivation for the current study.

Purpose of the Current Study

The purpose of the current study was to examine the roles of examinee age, task characteristics, and engagement in the context of gathering evidence about the validity of a technology-enhanced speaking test for YLLs. This involved evaluating a sample of *TOEFL Primary* Speaking test tasks and raters and confirming that they performed consistently across age groups. In addition, scores were demonstrated to be mostly attributable to examinees and not to tasks or ratings. Task characteristics that were hypothesized to support task engagement, along with their effects on measurement quality, were also evaluated. Lastly, the structural relationships between examinee age, task characteristics, engagement, and performance were explored. In addition to supporting a partial validity argument for the *TOEFL Primary* Speaking test and others like it, the anticipated results were expected to yield data to inform the practice of speaking test development while advancing a new approach of using acoustic measures to indicate engagement.

Research Questions

In order to gather evidence about the validity of a technology-enhanced speaking test for YLLs and to examine their engagement with the tasks themselves, the following research questions were posed.

1. How well do technology-enhanced speaking test tasks discriminate among young language learners of varying abilities, and is there any evidence of bias related to examinee age?

2. How consistent are scores across parallel tasks and ratings, and what is the optimal configuration of tasks and ratings for maximizing the dependability of test scores?
3. How do task characteristics, such as the presence or absence of topical choice, vocabulary support, novelty, and video animation, affect the quality of measurement (i.e., task difficulty, discrimination, and point-measure correlation)?
4. What is the nature of the relationships between such task characteristics, engagement, and performance, and do these relationships vary with examinee age?
5. How do young language learners describe speaking test tasks containing such task characteristics in terms of how engaging they are?

Definitions of Key Terms

The following terms are used throughout the study.

Young Language Learners (YLLs)

The term *young language learners* (YLLs) refers to learners of a second or foreign language—usually English—who are between the ages of approximately 8 and 13 years. Despite the large variety of terms that have been coined to refer to this population (Lara-Brady & Wendler, 2013), YLLs are markedly different from their adolescent or adult language-learning peers (Bailey, 2017). For example, the development of YLLs' language proficiency is typically viewed as a process of acquisition, not formal learning (Clark, 2000). This process of acquisition is closely linked to young learners' maturation and growth (Hasselgreen & Caudwell, 2016; Piaget, 1964), and it occurs in specific, predictable stages (Clark, 2000; Curtain & Dahlberg, 2010). Like first language learners, YLLs tend to develop oral skills first (Lopriore & Mihaljević Djigunović, 2011; Zangl, 2000) with a heavy emphasis on meaning and communicativeness (Cameron, 2003; Ellis, 2008).

Speaking Ability

Speaking ability involves having access to the "resources…to express, understand, dynamically co-construct, negotiate, and repair meanings, knowledge, and action, often in goal-oriented interaction" (Purpura, 2017, pp. 48–49). Typical language tasks that YLLs must engage in include narratives, descriptions, instructions, arguments, and opinions (McKay, 2006). To facilitate the assessment of YLL's speaking abilities, integration with other skills, such as listening and reading, with rich stimuli featuring language that approximates the language experienced in YLLs' everyday lives is recommended (Turkan & Adler, 2011). This does not mean, however, that the *content* of stimulus material must mirror the lives of YLLs. Language can also serve as a window to other possibilities (Bishop, 1990). YLLs readily incorporate new language and propositional content as they actively seek and negotiate meaning in their oral communications (Cameron, 2003; Oliver, 1998). Accordingly, this can inspire novel language production in the course of YLL performance.

Performance Tests

According to principles of evidence-centered design (Mislevy, Steinberg, & Almond, 2002), a well-designed test is one that gathers evidence that can support claims about examinees' knowledge, skills, abilities, or attributes. Since speaking ability, like other latent traits, is not directly observable, it can only be inferred from performances during assessment. Accordingly, a well-designed speaking test will feature sufficient tasks for eliciting performances as evidence for claims that can be made about examinees' underlying speaking abilities. While tasks may be different things to different people, it is commonly agreed that they involve the use of the target language to convey meaning and to carry out some authentic activity (Bygate, Skehan, & Swain, 2001). Performances on tasks that simulate real-life communicative activities are believed to be more predictive of language use in actual non-test contexts (Jones, 1985; McNamara, 1996; Lee,

2006). Descriptions of the *TOEFL Primary* Speaking tasks involved in the current study are provided in Chapter 3 and Appendix B.

Technology-Enhanced Items and Tasks

Technology-enhanced items allow for assessment possibilities beyond what traditional multiple-choice and constructed-response tasks can offer. In this paper, the term *technology-enhanced*, as it applies to items and tasks, subsumes a myriad of other commonly used terms, including *technology-enabled*, *innovative*, and *computer-based*. Technology-enhanced tasks can serve as alternative ways of gathering evidence about constructs, including hard-to-define constructs (Scalise, 2012). The view that technology-enhanced assessments are more engaging and therefore somehow more valid is widely held (Bryant, 2017). However, steps should still be taken to ensure the utility of such assessment tasks (Russell, 2016).

Parshall, Harmes, Davey, and Pashley (2010) delineated seven dimensions of technology-enhanced tasks that can reflect varying degrees of innovation: media inclusion, complexity, fidelity, assessment structure, response action, level of interactivity, and scoring methods. Media inclusion refers to the use of graphics, audio, and animation in the task. This dimension, especially as it relates to characteristics of stimuli, is a focus of the current study. Complexity describes how many and what kinds of elements the test taker must consider in order to successfully complete the task. Complexity has also been identified as a primary component of language task difficulty (e.g., Robinson, 2005; Robinson & Gilabert, 2007). Assessment structure refers to the variety of possible task formats and response types, including multiple-choice, hotspots, and drag-and-drop as variations of selected responses, potentially interacting with tools as variations in constructed responses (e.g., Oh, 2018), and variations in the number and connectivity of tasks (e.g., Banerjee, 2019). Response action refers to the actual physical ways of responding to tasks, such as clicking a mouse or speaking into a microphone. The level of interactivity describes how and how much a test or item responds to input from a test taker.

Another word for fidelity is authenticity, or how similar tasks reflect real-world situations in terms of the stimuli and expected responses. The scoring method refers to how responses are scored, including issues such as automated scoring and rubrics.

The use of rich multimedia content in task stimuli offers many affordances that affect the test-taking experience for YLLs (Cho & Getman, 2013). Colorful graphics, video animations, music, and spoken dialogue are believed to make test content more fun and engaging for YLLs (Hasselgreen, 2000, 2005; McKay, 2006; Taylor & Saville, 2002). By making test activities more playful and game-like, multimedia content may also mitigate the possible undesirable effects of cognitive and affective variables related to being a YLL, such as having a short attention span and language test anxiety, respectively (Bailey, 2017).

Closely related to featuring multimedia content is the ability to create virtual worlds that deemphasize the effects of YLLs' experiential limitations. By contextualizing activities in age-appropriate scenarios, such as a trip to outer space or an underwater exploration, speaking assessments may address topics and situations that would be difficult to emulate in a face-to-face test format. The expanded range of situations presented in the speaking assessment makes it possible to elicit more diverse samples of language use. For example, fantastical situations that would not occur in real life can facilitate novel language production by test takers instead of more routine expressions.

Virtual worlds and scenarios can also support a diverse cast of characters. Characters can create richer opportunities for purposeful and meaningful oral language production than would typically be possible in an interview (Bailey, 2017). For example, explaining a sequence of events from the demonstration of an activity to a virtual peer who missed seeing the demonstration is a more authentic speech activity for young learners than simply retelling a story to a test administrator would be. Multiple characters allow for a wide range of interactional patterns within an assessment.

The ability to employ multimedia to create scenarios complete with a cast of characters in a speaking assessment can also allow for scaffolding to be provided to assist test takers to achieve what they may not be able to achieve independently. For example, introducing relevant vocabulary words aurally, pictorially, and textually at the beginning of some tasks may support YLLs to demonstrate their abilities to actually use the language instead of whether they know a few key words. Likewise, possible oral responses to tasks can be modeled by characters in a scenario to help clarify the expectations of a task. Such supports can occur seamlessly in a computer-delivered assessment.

Computer delivery also allows for the digital recording of responses. When responses are unobtrusively captured, test takers can perform freely without concerns related to shyness or trying to please adult interviewers. In addition to reducing the impact of affective variables, recorded responses can later be distributed and scored. A distributed scoring model has many benefits. One of these benefits is improved score quality. For example, instead of a single interviewer assigning scores to all the responses from an individual test taker, responses from the test taker can readily be distributed to and scored by several qualified raters. This would decrease the influence of a single rater on the test taker's total score, making scores more dependable. This approach to collecting and scoring responses also addresses local concerns that there may not be enough qualified teachers to administer and score speaking tests in a face-to-face format (Cho et al., 2016, 2017).

Task Characteristics

All tasks, including technology-enhanced ones, have certain characteristics. For example, task characteristics can vary in terms of participants, purposes, form and content, tone, language, norms of interaction, genre, and problem to be addressed (Douglas, 2000). Task characteristics examined in the current study are characteristics of the input. "Input consists of the material contained in a given test task…which the test takers or language users are expected to process in

some way and to which they are expected to respond" (Bachman & Palmer, 1996, p. 52). Taxonomies of task characteristics that influence the relative difficulty of tasks have been studied (e.g., Robinson, 2005; Robinson & Gilabert, 2007). However, taxonomies of task characteristics that may support engagement have not received as much attention in the literature. Thus, several categories of task characteristics suspected to support engagement are described in Chapter 2. Some of these task characteristics are the focus of the current investigation, and they are briefly introduced here.

Topical choice. Choice is a characteristic of a task that gives test takers options about which topic to discuss in their responses. Sometimes a set of options can be presented as suggested topics for the test taker to choose from, and other times the choice can be open to any topic relevant to a task. Choices, even superficial ones, have long been recognized as supporting learner engagement (e.g., Cordova & Lepper, 1996).

Vocabulary support. This task characteristic generally occurs at the beginning of a task. By seeing pictures, reading text, and hearing the pronunciation of select lexical items that might be useful to completing a task, examinees are enabled to show what they can actually do with the language instead of being impeded by not knowing a few key words. The words presented are optional in that their use is not required in order to complete or receive the highest scores on the task.

Novelty. Novelty refers to the presentation of unexpected, fantastical situations that would not be expected to occur in real life, such as a bus being driven by a giraffe. Novelty and fantasy have long been thought to stimulate interest in students (e.g., Bergin, 1999; Malone, 1981).

Video animation. Animations, commonly referred to as cartoons, are motion pictures made from drawings that depict some sort of action. Sound or music often accompanies video

animations. Such video animations can generally provide richer context to assessment tasks (Parshall et al., 2010).

Engagement

Engagement is "the level of involvement or connection between person and activity" (Ainley, 2012, p. 285). In the current study, the term refers to examinee engagement with a language test task. Presumably, the more engaging a task is, the better responses to it will reflect latent abilities like speaking ability. However, there is a paucity of research on engagement at the local level of the task. The bulk of the literature focuses on more general school engagement (e.g., Lawson & Lawson, 2013) and engagement in the workplace (e.g., Macey & Schneider, 2008). A better understanding about the nature of task engagement is clearly needed.

Engagement and motivation are related constructs (Appleton, Christenson, Kim, & Reschly, 2006), which has led to some significant overlap in the use of the terms in the literature. While motivation has been a long-standing subject of interest among language researchers (e.g., Clément, 1986; Dörnyei, 2005; Gardner, 1985; Gardner & Lambert, 1972), motivation is often viewed as preceding engagement (Meltzer & Hamann, 2004). This view is adopted in the current study. The distinction between motivation and engagement is that motivation is the underlying psychological *trait* that turns into the *state* of engagement through "energized, directed, and sustained action" (Skinner, Kindermann, Connell, & Wellborn, 2009, p. 225).

A process model of engagement is useful when describing engagement at the task level. Dörnyei (2002) described a process model that distinguishes three stages of an activity: the pre-actional stage, the actional stage, and the post-actional stage. The pre-actional stage involves underlying motivations, including the setting of goals and intentions before beginning the activity. The actional stage describes acts of engagement that involve performing subtasks, appraising one's performance, and regulating one's attention during the activity. The post-actional stage involves a retrospective evaluation of the activity, and it may even result in a

feedback loop that influences the motivation to engage in similar activities in the future. For example, among adults, engagement with one language learning task was related to being motivated to engage in future tasks (Fryer, Ainley, & Thompson, 2016). Lawson and Lawson (2013) expanded upon this process model of engagement by breaking up the pre-actional stage into personal attributes and task characteristics that both contribute to engagement.

The process model of engagement (see Figure 1.1) shows how personal attributes like motivation interact with task characteristics to result in acts of engagement, such as responding to a speaking task. While there has been some research on the variable nature of YLL motivation and attitudes (e.g., Carreira, 2006; Heining-Boynton & Haitema, 2007; Mihaljević Djigunović & Krevelj, 2009; Nikolov, 1999), there is little that can be done from a test development point of view to influence such personal attributes except, perhaps, trying to make sure that the consequences of engaging with a test are positive and support motivations for further language use and learning. Practically speaking, in the context of a language assessment, only the task characteristics can be readily manipulated by test developers. In Chapter 2, a taxonomy of task characteristics that may support engagement is proposed and discussed.

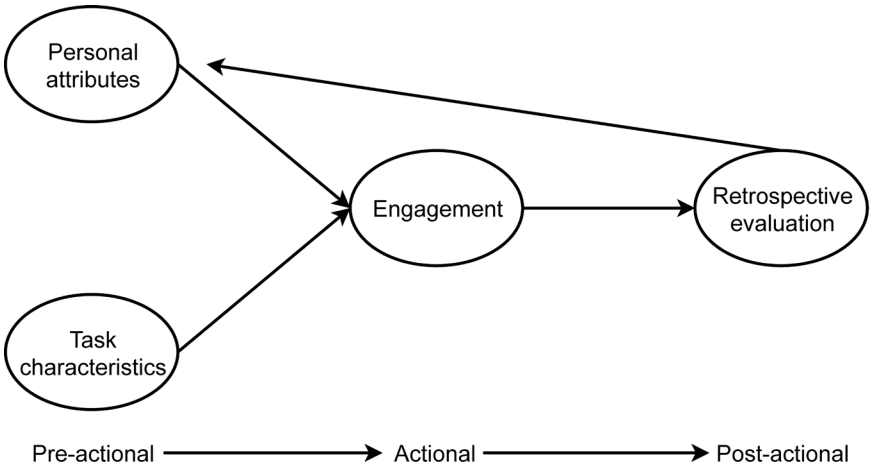


Figure 1.1. Process model of engagement.

Validity

Notions of construct validity, which assert that a test actually measures what it claims to measure (e.g., Cronbach & Meehl, 1955), have significantly expanded. Messick (1989) described validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). In other words, the interpretation and uses of tests should be supported by both evidence and potential consequences. However, under the framework of this unified theory of validity, it could be difficult to know what kind and how much evidential support is really needed (Shepard, 1993).

In response, Kane (1992, 2006, 2013) proposed an argument-based approach to test validity that involves making an interpretive argument that links test performance to test use using a chain of inferences. Figure 1.2 graphically shows an example of an interpretive argument for a language test. The interpretive argument is strung together by a chain of inferences that includes domain definition, evaluation, generalization, explanation, extrapolation, decision, and consequence (Chapelle, Enright, & Jamieson, 2008, 2010; Knoch & Chapelle, 2018). Each inference is associated with a claim that can be supported or refuted by evidence. As evidence is gathered to support warrants and assumptions that underlie each claim and inference in the chain, the justification for test use becomes stronger.

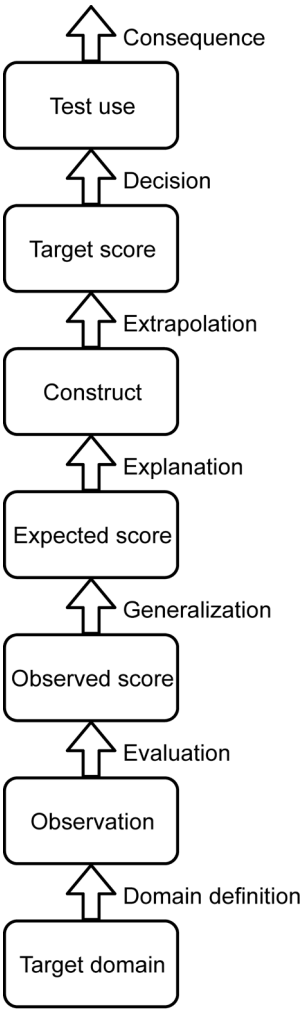


Figure 1.2. Validity argument for a language test. Adapted from *Building a Validity Argument for the Test of English as a Foreign Language* (p. 15), by C. A. Chapelle, M. K. Enright, and J. M. Jamieson, 2008, New York, NY: Routledge. Copyright 2008 by the Taylor & Francis Group.

The inference of domain definition is needed to connect the target domain to an observation of a language test performance. The claim associated to the domain definition is that performances on the test tasks reveal the language abilities required in the target domain. Evidence that assessment tasks can be developed to elicit the targeted language abilities is one way this claim can be supported. In the context of a speaking test for YLLs, evidence already exists to support the claim (e.g., Cho et al., 2016, 2017; Turkan & Adler, 2011).

Between the observation and the observed score is the evaluation inference. The associated claim is that observations of language performances result in scores that reflect the targeted language abilities. Evidence that would support this claim includes demonstrating that the rubrics are appropriate for providing information about the targeted speaking abilities and that raters are free of bias for or against any particular group. The current study gathered evidence to back up the claim related to the evaluation inference.

A generalization inference comes between the observed score and the expected score. The claim is that observed scores are estimates of expected scores across parallel versions of tasks and test forms and across raters. Relevant evidence would include showing that tasks and observations are representative of the universe of admissible tasks and observations and that the configuration of tasks and observations is sufficient to control sampling error (Kane, 2006). The current study also gathered evidence to back up the claim associated to the generalization inference.

Between the expected score and the construct is the explanation inference. The associated claim is that scores are attributable to a defined construct as specified by a theoretical model of language proficiency (Chapelle, 1998). Relevant evidence would show that scores are indicators of underlying language abilities as theorized by the model. In addition, task difficulty should be systematically influenced by task characteristics in accordance with the model. Lastly, tasks and task characteristics should not function differentially across groups except as specified by the model. Some of these assumptions and the evidence backing them were also investigated in the current study.

The last three inferences are extrapolation, decision, and consequence. Extrapolation connects the construct to the target score. The claim associated with extrapolation is that the construct assessed relates to the quality of performance required in the target language use domain. Although evidence could include correlating test performance with observations of

classroom performance or performance on other valid measures of proficiency, aligning scores to a framework such as the Common European Framework of Reference is common practice (e.g., Papp, 2018). Between the target score and test use is the decision inference. The associated claim is that decisions based on the test scores are appropriate. This could involve gathering evidence for the effectiveness of the score scale for differentiating test takers. Beyond test use is the inference of consequence. The associated claim is that the consequences are beneficial to users, and evidence could be gathered to back or refute this claim. For example, the effects of testing on YLLs themselves in terms of opportunity, their educational experiences in terms of washback, and their communities' institutions in terms of policy are all relevant sources of evidence for this claim (Rixon, 2018). However, extrapolation, decision, and consequence inferences are beyond the scope of the current study.

Significance of the Study

The current study is expected to contribute to the field in several ways. Since very little has been published about the validity of assessments for YLLs (Butler, 2017a), one goal of the current study was to help fill that gap in the literature. For instance, empirical evidence was gathered about claims associated to the evaluation, generalization, and explanation inferences of a validity argument for a technology-enhanced speaking test for YLLs. Of particular relevance to this population is whether variations in examinee age pose a threat to claims about the usefulness of scores. The role of examinee age in task performance, rater scores, and the meaning of those scores had not previously been well studied.

The nature of task engagement was also a focus of the current study. In particular, the influence of certain task characteristics on engagement and performance was investigated. In addition to proposing a taxonomy of task characteristics that may support engagement, the study reports how acoustic and prosodic features were extracted from spoken responses and evaluated

as potential indicators of engagement. Findings about the effects of topical choice, vocabulary support, novelty, and video animation on engagement were then confirmed qualitatively through retrospective verbal reports. The nature of engagement in young learner assessments had also not previously received much attention in the literature.

In addition to applying a validity argument to a speaking test for young learners and exploring the role of task engagement in a language test, the current study suggests that engagement should be treated as an essential component in theoretical models of language proficiency. Examinee age, on the other hand, appeared to have little interaction with tasks and raters and little effect on task engagement and performance. These findings suggest that the *TOEFL Primary* Speaking test may be valid throughout the targeted age range. Other practical implications related to task design are discussed in Chapter 5.

Summary

In response to the demand for valid speaking assessments for YLLs, the current study aimed to explore the nature of speaking performance and task engagement among YLLs from a test development point of view. The intention was to evaluate the influences of examinee age and task characteristics that may support engagement on performance within the context of gathering evidence for a validity argument. Also, the influences of these task characteristics on potential acoustic indicators of task engagement as well as YLLs' retrospective verbal reports were investigated. This study was carried out to help inform and improve future technology-enhanced test task design for YLLs in general and to provide evidence regarding the validity of the *TOEFL Primary* Speaking test in particular.

Chapter II

REVIEW OF THE LITERATURE

Chapter 2 examines prior research related to the current study, and it identifies gaps that the current study is intended to help fill. This literature review covers three broad areas. First, attributes of young learners of a second or foreign language that may affect their speaking test performances are described. Second, theoretical and practical considerations related to speaking assessments and their validity are examined. Then, the nature of task engagement is explored. Where these three spheres overlap sets the backdrop for the current study.

Considerations Related to Young Learner Age

Middle childhood (i.e., between the ages of approximately 8 and 13 years) is a time of intense experiential, cognitive, and psychosocial growth (Hasselgreen & Caudwell, 2016). It is self-evident that young learners, because of their limited life experiences, may be more sensitive to variations in assessment task content than older learners. Even the experience of the language classroom can vary considerably, from awareness-raising programs to language-focused ones and from content-based curricula to full immersion (Inbar-Lourie & Shohamy, 2009). Cultural variations can also add to the complexity of appropriately assessing ELLs (Pitoniak et al., 2009). The limited life experiences of children may also heighten the role of culture in language performance (Johnstone, 2000). The effects of these limitations in experience on language test performance may be further compounded by YLLs' still-developing cognitive capacities, which affects how YLLs interact with task content—especially multimedia content—and by psychosocial variations.

Cognitive Development

Cognitive development plays a substantial role in language acquisition among YLLs. A critical or sensitive period in which YLLs have a particular advantage in language learning over their adolescent and adult counterparts has been suggested by the literature (DeKeyser, 2000; Hyltenstam & Abrahamsson, 2003; Johnson & Newport, 1989), especially in the areas of pronunciation (Bongaerts, Van Summeren, Planken, & Schils, 1997) and phonemic discrimination (Larson-Hall, 2008). However, these findings are not without their critics (e.g., Bialystok, 2002; Marinova-Todd, Marshall, & Snow, 2000; Moyer, 1999). Today it is generally accepted that while the level of ultimate attainment may be higher for learners who start at an early age, older learners, with their more developed cognitive capacities, can often make faster learning gains with appropriate exposure and instruction (Muñoz, 2009).

Unlike the influences of cognitive development on language acquisition, the role of cognitive development in language test performance is not as well studied. However, from a processing perspective (see Vanpatten & Cadierno, 1993), it is likely that maturational constraints on cognitive capacities would impact the quality of output, especially with productive language tasks like speaking. Therefore, when language assessments are developed, considerations should be made for the age of targeted test takers. After all, plenty of evidence exists that shows measurable differences in cognitive abilities that can be attributed to age.

One of the most salient examples of the rapidly expanding cognitive abilities of YLLs (and their native-speaking peers) is working memory capacity. The “most widely quoted model” of working memory (Field, 2018) was originally proposed by Baddeley and Hitch (1974) and later expanded by Baddeley (2000, 2003, 2015). In this model, working memory is composed of four distinct components (see Figure 2.1). A central executive controls attentional resources. Then, there are separate storage systems for auditory and visual information that became known as the phonological loop and the visuospatial sketchpad, respectively. The fourth component, the

episodic buffer, is theorized to be able to integrate temporary representations from other systems and interface with long-term memory.

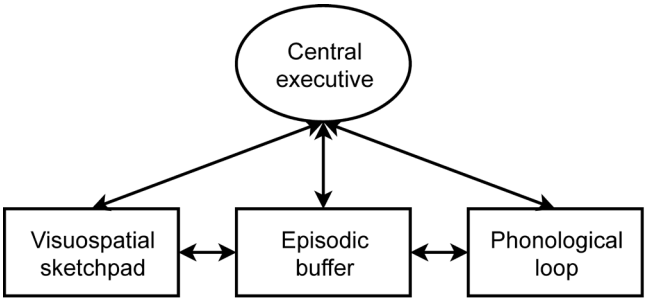


Figure 2.1. Model of working memory. Adapted from "The Episodic Buffer: A New Component of Working Memory?" by A. D. Baddeley, 2000, *Trends in Cognitive Sciences*, *4*(11), p. 421. Copyright 2000 by Elsevier Science.

Several studies have demonstrated measurable differences in the capacities of the visuospatial sketchpad and the phonological loop of children at different ages. For example, Anderson and Lajoie (1996) administered several different cognitive tasks that measured verbal memory and visuospatial memory to 376 children between the ages of 7 and 13 years of age. The results revealed consistent and predictable improvements in performance across all measures that corresponded with increases in age. Likewise, Gathercole, Pickering, Ambridge, and Wearing (2004) investigated the performance of approximately 600 children on tasks intended to measure verbal storage alone, complex memory span (i.e., being able to process and store information simultaneously), and visuospatial memory. In addition to recording a sizeable expansion across all measures between the ages of 4 and 15, the researchers examined the factor structure of working memory, which they concluded was largely invariant across age groups. Though these studies involved native speakers, the same patterns of cognitive development are expected in YLLs.

The expanding capacity of phonological working memory may have unique implications for YLLs. For example, phonological working memory appears to significantly predict second

language achievement during the school years. A series of longitudinal studies of Finnish children between the ages of 7 and 13 revealed that eventual performances on English listening, reading, and writing tests were highly correlated with measures of phonological working memory (Dufva & Voeten, 1999; Service, 1992; Service & Kohonen, 1995). In another example, French (2003) reported on a study of 54 eleven-year-old YLLs whose performances on an English vocabulary test were strongly predicted by measures of phonological working memory, especially among YLLs of lower ability. In a follow-up study of 104 YLLs of the same age, French and O'Brien (2008) found that phonological working memory also predicted performance on a test of morphosyntactic structures. While the evidence suggests that phonological working memory supports language learning and performance in reading, listening, and writing, the effect of cognitive capacities that develop with age on YLL speaking performance is not so clear. The current study investigated interactions between age and speaking performance.

Interactions with Media Content

Cognitive variations due to age may also affect how YLLs interact with the stimulus content of tasks itself. According to Mayer's (2005) cognitive theory of multimodal learning, humans have separate channels for processing graphical and linguistic information, but these channels have limited capacities. For adult learners, presenting content in multiple modes (e.g., pictorially and audibly) leads to a split-attention effect (Mayer & Moreno, 1998), which seems to enhance uptake with both channels, compared to presenting content in one mode only (e.g., Al-Seghayer, 2001; Chun & Plass, 1996; Duquette et al., 1998; Jones & Plass, 2002; Mohsen & Balakumar, 2011). However, these findings may not apply to YLLs, whose cognitive development likely affects the capacities of the two separate channels for processing information.

One study that did involve YLLs exposed to different kinds of multimedia was conducted by Acha (2009). She reported on a study of 135 third- and fourth-grade children in Spain who read a short story in English on a computer. Twelve previously unknown words were presented

with glosses of three types: written translations, pictures, and a combination of the two. Later recall of new vocabulary meanings was better for children who received only textual annotations than for children who received both pictorial and textual annotations or pictorial annotations alone. This suggests that presenting redundant information through multiple channels may result in some cognitive bottlenecking for YLLs (Sweller, 1994).

On the other hand, Getman, Cho, and Luce (2016) investigated the effects of presenting multiple-choice listening items under two different conditions: with the options presented only aurally and with them presented both aurally and in printed text form in a test book. Participants were 747 YLLs between 7 and 15 years of age from Mongolia, Columbia, and Brazil. The results indicated that there was not a significant difference in task performance between the two conditions, regardless of age or reading ability level. YLLs also reported that they preferred the items presented both aurally and as printed text. These findings suggest that the redundancy of presentations, in this case, was not too taxing cognitively for YLLs. Work still remains to better understand how multimedia stimuli interact with age-related variations in cognitive capacity and how these interactions might affect YLL speaking performances. These issues are also addressed by the current study.

Psychosocial Variation

In addition to cognitive factors, psychosocial factors like affect and interpersonal orientation influence the language use of YLLs (Mihaljević Djigunović, 2009). For example, YLLs tend to have a "heightened sensitivity to praise, criticism, and approval" (McKay, 2006, p. 14). For this reason, in terms of speaking assessments, an interlocutor can play a substantial role in YLL performance. Kondo-Brown (2004) conducted a study of 30 fourth-grade students who were learning Japanese as a foreign language. During a test of oral proficiency, interviewer support in the form of explicit correction, implicit correction, repeating the question, or making clarification requests tended to result in higher ratings than with the absence of such support. Of

course, this variability due to the actions of an interlocutor is usually not desirable, especially in the context of standardized assessments.

Shyness among YLLs is another problem with oral proficiency interviews, especially when they are conducted by a stranger (Cho et al., 2016, 2017). Shyness from such novel social or evaluative situations is closely related to anxiety (Coplan, Prakash, O'Neil, & Armer, 2004). Standardized tests are already known to induce anxiety among young learners (Segool, Carlson, Goforth, von der Embse, & Barterian, 2013). YLLs in Taipei reported that tests and speaking in front of others are two of the most anxiety-inducing situations they face (Chan & Wu, 2004). Since anxiety can impair cognitive functioning (Eysenck, Derakshan, Santos, & Calvo, 2007; Owens, Stevenson, Norgate, & Hadwin, 2008), shyness would also likely impair performance on oral interviews.

To test this claim, Crozier and Hostettler (2003) administered vocabulary tests in three different conditions—in face-to-face oral interviews, in face-to-face written interviews, and in a group setting—to 240 Year 5 pupils in Britain. They discovered that shyer children performed poorly in the face-to-face formats but did as well as their peers in the group settings. This suggests that the face-to-face interviews introduced construct-irrelevant variance to their vocabulary scores. Though the participants in this study were native speakers, it is reasonable to assume that these findings would generalize to YLLs taking speaking tests. For this reason, computerized tests that record responses for later scoring may be the preferred format over interviews for assessing YLL speaking ability (Cho et al., 2016, 2017). Such findings informed the development of the instrument in the current study.

Experiential constraints, developing cognitive capacities, and affective and interpersonal factors that are all related to the age of YLLs likely need to be considered when developing language assessments for them. Therefore, examinee age also deserves special consideration when investigating the validity of such language assessments. The next section reviews literature

about the validity of speaking tests and highlights areas where the role of examinee age may be particularly relevant.

The Validity of Speaking Tests

Contemporary approaches to test validity research involve gathering evidence to support or refute claims that underlie an interpretive argument for a test's validity (Kane, 1992, 2006, 2013). The interpretive argument for a language test connects language test performance to test use with a chain of inferences (see Chapelle et al., 2008, 2010; Knoch & Chapelle, 2018). Table 2.1 lists inferences and the associated claims that support the validity of a speaking test. The current study focused on gathering evidence related to the evaluation, generalization, and explanation inferences for a technology-enhanced speaking test for YLLs.

Table 2.1

Inferences for the Validity of a Speaking Test and Their Associated Claims

Inference	Claim
Domain definition	Observations of performances reveal knowledge, skills, and abilities relevant to speaking in the target language use domain.
Evaluation	Observations are evaluated using procedures that provide observed scores with intended characteristics.
Generalization	Observed scores are estimates of expected scores over relevant parallel tasks, ratings, and test forms.
Explanation	Scores are attributable to a construct of speaking proficiency.
Extrapolation	The assessed construct sufficiently accounts for the quality of speaking performances in the target language use domain.
Decision	Decisions made based on the estimates of the quality of the performance are appropriate and well communicated.
Consequence	Test consequences are beneficial to users.

Note. Adapted from "Validation of Rating Processes Within an Argument-Based Framework," by U. Knoch and C. A. Chapelle, 2018, *Language Testing*, *35*(4), p. 482. Copyright 2017 by Sage Publishing.

In this section, relevant literature regarding warrants and assumptions underlying each of the three inferences and their claims as they relate to speaking tests is presented. First, in support of the evaluation inference, the effectiveness of scoring systems, which includes issues of rubric functioning and rater bias, is discussed. Second, a brief overview of studies involving the application of generalizability theory to speaking test scores is presented. Lastly, in support of the explanation inference, relevant theoretical models of speaking performance are described.

The Evaluation Inference

The evaluation claim states that scores from observed performances have the intended characteristics. In other words, they effectively discriminate among examinees according to their speaking abilities. Warrants underlying this claim state that the scale properties are as intended by test developers and that raters rate reliably at the task level (Knoch & Chapelle, 2018). Assumptions to be investigated that are relevant to these warrants include having tasks and rating scales that are appropriate for providing information about targeted speaking abilities and having raters who can consistently apply the rating scale and who are free of bias for or against any particular group, such as one based on age. However, it is important to note that neither raters nor rubrics exist in a vacuum. Raters apply the rubrics to responses, and there may be a myriad of ways to support or augment this relationship. For example, benchmark or anchor responses, which exemplify score bands on a rubric and help to bring rubric descriptors to life, may affect the performances of both the rubric and the rater. The same goes for rater training, active monitoring and feedback by scoring leadership, ongoing calibration tests at the start of each scoring session, and many other policies and procedures used to help ensure score quality. Therefore, any evidence gathered about rubrics or raters really supports (or refutes) assumptions about the effectiveness of the scoring *system* instead of its individual components in isolation.

With that clarification in mind, many studies about the functioning of speaking tasks and scoring systems have been conducted using many-facet Rasch analyses. Rasch analysis involves

modeling probabilities of measurements as logistical functions of multiple parameters, or facets, such as examinee ability, task difficulty, and rater severity (see Bond & Fox, 2015). Unlike 2-parameter item response theory (Birnbaum, 1968), Rasch measurement is prescriptive, meaning that data are expected to fit theoretical models, and it is consistent with principles of fundamental measurement (Wright, 1999). For these reasons, Rasch analysis can be a good choice for evaluating the functioning of tasks, rating scales (Linacre, 2002), and raters (McNamara, 1996; Myford & Wolfe, 2003). It can also be useful for investigating potential systematic interaction effects between facets, such as those between raters and examinees (e.g., Johnson & Lim, 2009; Kim, 2011; Kondo-Brown, 2002; Kozaki, 2004; Lynch & McNamara, 1998; Schaefer, 2008; Winke et al., 2011, 2013).

One such study was performed by Winke, Gass, and Myford (2011, 2013). They were interested in seeing if there was any evidence of rater bias related to accent familiarity. The 72 adult test takers, whose first languages were Spanish, Chinese, or Korean, each responded to six different *TOEFL*® iBT speaking tasks on a computer. The responses were then scored by raters with experience learning Spanish, Chinese, or Korean as a second language. Data were fit to a Rasch model, indicating that tasks and raters functioned as expected in discriminating among examinees according to ability. However, bias analyses revealed some evidence of interaction between examinees grouped by their first languages and raters grouped by the second languages they had learned. Specifically, examinees in the Spanish and Chinese native language groups were rated more leniently by raters with experience learning the same language, suggesting that rater familiarity with examinee accents may be a factor in rater variation. Although the study did provide some evidence that evaluations of speaking performances to computerized tasks resulted in scores with expected characteristics, the study was conducted on adults, so it is unclear how well this finding would generalize to a YLL population. In addition, although investigating accent familiarity as a possible factor in rater variation may apply to a speaking test for YLLs, the

possibility of examinee age being a factor in the functioning of raters and tasks seems to be a more pressing investigative need.

The role of examinee age in constructed-response assessments has been largely neglected in the literature, but it is particularly relevant to assessing YLL speaking. Examinee age could be a variable that has unexpected effects on ratings of spoken responses. For example, anecdotally, raters have commented about how "cute" some responses from the youngest learners are, but it is unknown if those perceptions have any effect on rater severity. In addition, examinee age is correlated with experience, cognitive development, and affective and interpersonal variations, as discussed above. Examinee age has the potential to interact with task content, thus affecting scores. Clearly, the meaningfulness of scores that are assigned to YLLs' spoken performances needs verification. The evaluation inference that contributes to the overall argument for the validity of a speaking test for YLLs demands that evidence be gathered to support assumptions about the effectiveness of tasks and the scoring system to discriminate among examinees by their varying abilities and not by age. This evidence was gathered as part of the current study.

The Generalization Inference

After the evaluation inference comes the generalization inference. The claim of the generalization inference is that observed scores are estimates of expected scores across parallel ratings, tasks, and test forms. Evidence that examinee effects, not rater or task effects, are the primary contributors to score variability would support this claim. The claim could be further supported by evidence that the configuration of tasks and ratings is appropriate to control sampling error. Such evidence is typically gathered through the applications of generalizability theory (Brennan, 2001a; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991).

"Generalizability theory (G-theory) is a measurement model that enables us to investigate the relative effects of multiple sources of variance in test scores" (Bachman, 2004, p. 176). A test

score is conceived as a sample from a universe of admissible observations. This universe is composed of all the observations that could result from possible combinations of facets, such as examinees, tasks, and raters. Disentangling and estimating sources of score variance is done through a generalizability study.

After determining the variance components with a generalizability study, decision studies can be undertaken to test different configurations of facets in the measurement design to maximize score dependability. For example, Lee (2005, 2006) reported on a study of the speaking section of the *TOEFL* test. He determined that the greatest source of score variance was attributable to examinees, as expected. He then conducted a series of decision studies to investigate the effects of varying the number of tasks and raters on overall score dependability. He concluded that overall dependability of measurement was maximized by increasing the number of tasks rather than by increasing the number of raters rating each task. These results are slightly different from those from a study of Australia's Access test, for which decision studies showed a greater benefit in having more raters than in having more tasks (Lynch & McNamara, 1998). The differences in the findings between these two studies may be due to variations in the nature of the tasks and the scoring systems involved. The current study also employed decision studies to identify the optimal configuration of tasks and ratings to maximize the dependability of scores across parallel test forms on a test for YLLs.

The Explanation Inference

In addition to the evaluation of performances and the generalization of scores, an explanation about the meaning of scores is also a component in an interpretive argument for a test's validity. The meaning of scores relates to the constructs underlying the language test, which must be derived from a theoretical model of language ability (Chapelle, 1998). Therefore, the question of what it means to know a language (Spolsky, 1973) is an essential one for test development. Since the 1960s, many models of language ability have been proposed and

discussed, and each in turn has reflected some sort of expansion to accommodate evolving understandings of language proficiency. In order to understand how scores relate to underlying constructs, a brief summary of some of the major theoretical developments in the history of the construct is presented.

The earliest models of language ability reflected a structuralist point of view, which looks at language ability as being made up of component parts. Lado (1961) proposed a "skills-and-elements" model of language proficiency composed of three discrete elements (phonology, structure, and the lexicon) in each of the four skills (listening, speaking, reading, and writing). After measuring these discrete components through a variety of task types, it was presumed that the sum of measurements would reflect one's language ability.

Carroll (1961, 1968) built on the "skills-and-elements" model. Figure 2.2 shows how he distinguished morphology and syntax as distinct structural elements, but, more importantly, he pointed out the need to consider another dimension—the "total communicative effectiveness of an utterance" (Carroll, 1961, p. 37). Eventually, this notion of communicate effectiveness grew into an argument for testing "integrative skills," or multiple skills being used in tandem, in addition to testing the other discrete elements (Carroll, 1968). This is one of the first suggestions that language is somehow more than the sum of its parts.

	Phonology / orthography	Morphology	Syntax	Lexicon	Integrative skills
Listening					
Speaking					
Reading					
Writing					

Figure 2.2. "Skills-and-elements" model of language ability.

Oller (1979) abandoned preceding structuralist notions of language ability by proposing the unitary trait hypothesis, which holds that language performance involving different skills and different contexts draws on the same set of linguistic resources. His model of language proficiency was based largely on what he called a "pragmatic expectancy grammar," or an

expanded notion of grammar that included the ability for linguistic elements to be mapped onto extralinguistic contexts. Accordingly, a language user's internalized grammars reflect a unitary competence, or one's general language proficiency, that would be expressed through various components that are really inseparable (Oller, 1983).

Subsequent evidence has pointed to multicomponential models of communicative language ability (e.g., Bachman & Palmer, 1982; Sang, Schmitz, Volmer, Baumert, & Roeder, 1986; Song, 2008), but there have been differing views about which factors actually play roles in language proficiency. For instance, following an expansion of Hymes' (1972) notion of communicative competence, which includes "ability for use" in addition to grammatical knowledge, Canale and Swain (1980) proposed a multicomponential model of communicative competence that includes grammatical, sociolinguistic, and strategic competencies—the latter reflecting a compensatory mechanism, or a set of repair strategies, that could be applied in the event of communication breakdown. Canale (1983) further identified discourse competence, or being able to observe rules of discourse, including elements of coherence and cohesion, as another component. According to this model, these four competencies would be invoked by speaking test tasks (see Figure 2.3).

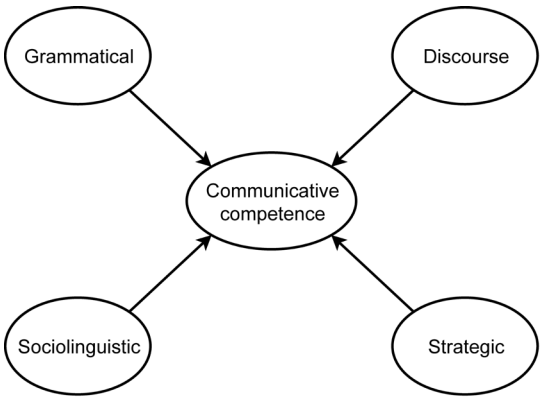


Figure 2.3. Model of communicative competence (Canale, 1983).

Expanding notions of strategic competence to include metacognitive strategies applied throughout a communicative task, from planning to execution, Bachman's (1990) and Bachman

and Palmer's (1996, 2010) model of communicative language ability has become one of the most influential models to date. Its sophistication lies in the inclusion of several components in addition to language ability (see Figure 2.4). Language ability consists of language knowledge (which includes grammatical knowledge, textual knowledge, functional knowledge, and sociolinguistic knowledge) and strategic competence. Along with language ability, topical knowledge and personal attributes, like age and background, are mediated by affect, strategic competence, and cognitive strategies. These components then all interact with the characteristics of the task. Task characteristics include the characteristics of the setting, the characteristics of the assessment rubric, the characteristics of the input, the characteristics of the expected response, and the relationship between the input and the expected response. Because of the "interactiveness" between attributes of the individual and task components, this model describes how performances can systematically vary.

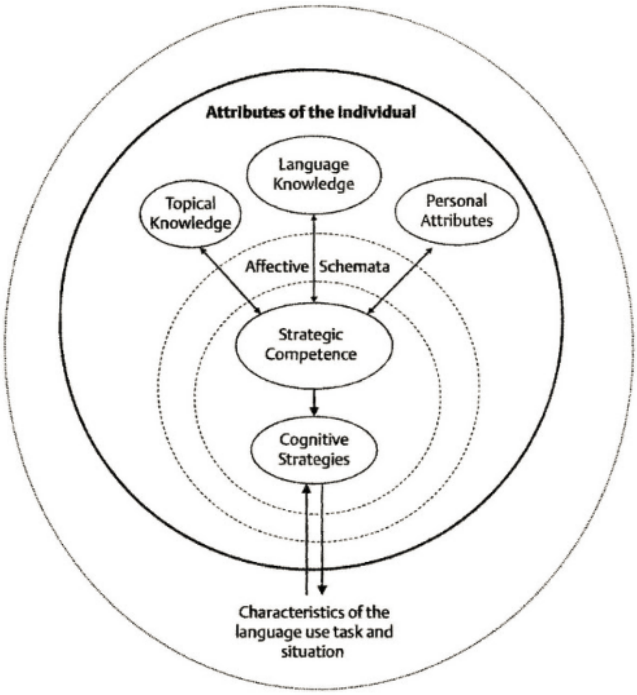


Figure 2.4. Model of communicative language ability. From *Language Assessment in Practice* (p. 36), by L. F. Bachman and A. S. Palmer, 2010, Oxford, UK: Oxford University Press.

Interactionalist approaches to modeling language ability expand upon this notion of interactiveness by distinguishing the context from the task and highlighting how both interact with each other and the language user. According to Chapelle (1998), "from an interactionalist perspective, performance [on a task] is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts" (p. 43). In other words, in addition to attributes of the language user and the characteristics of the task, contextual features also need to be specified by the model in order to be able to generalize performances from one context to another (see Figure 2.5).

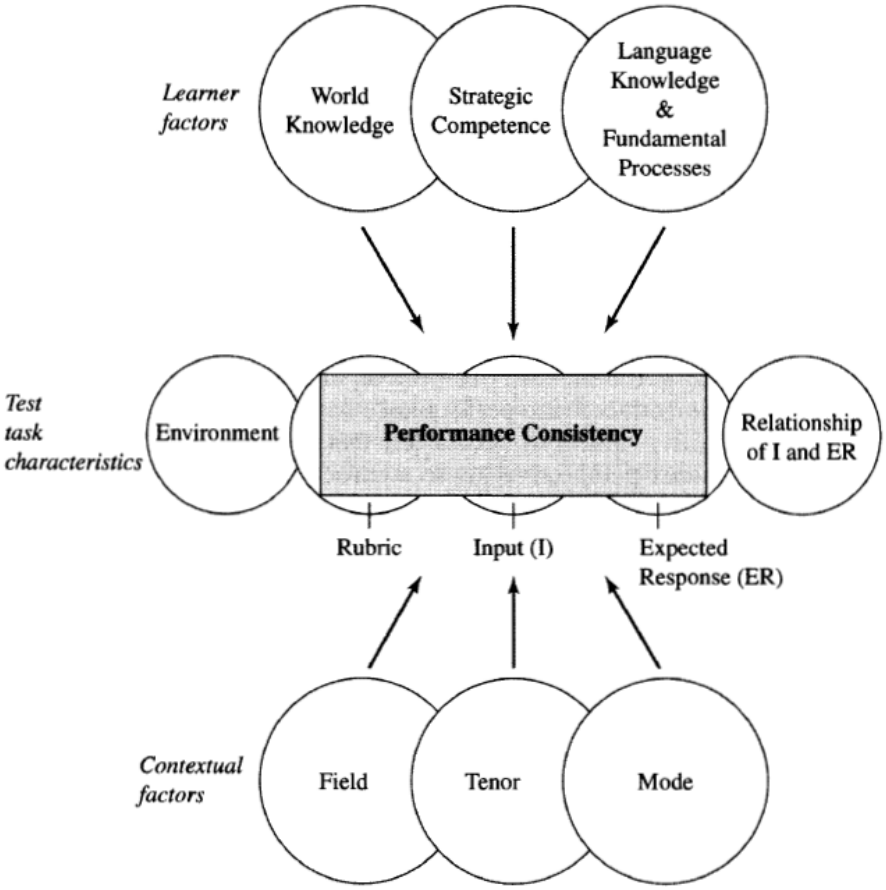


Figure 2.5. Interactionalist model of language performance. From "Construct Definition and Validity Inquiry in SLA Research," by C. A. Chapelle, in L. F. Bachman and A. D. Cohen (Eds), *Interfaces Between Second Language Acquisition and Language Testing Research* (p. 52), 1998, Cambridge, UK: Cambridge University Press. Copyright 1998 by Cambridge University Press.

One way to categorize contextual features that would be relevant to speaking performance is in terms of field, tenor, and mode (Halliday & Hasan, 1989). Field refers to the topic, the setting, and activities relevant to the situation. Douglas (2000) argued that language use is field dependent, so speaking tests "must engage test takers in a task in which both language ability and knowledge of the field interact with the test content in a way which is similar to the target language use situation" (p. 7). Tenor refers to other participants in a situation, including their roles and relationships. He and Young (1998) pointed out how interlocutors become relevant to a model of speaking proficiency, as interaction with them can lead to the co-construction of language and meaning. The mode is what part language plays in a situation, including the channel, genre, and meaning relations of the situated language.

Building on the significance of context in models of language performance, Purpura (2004, 2017) defined language proficiency as having the language, topical, sociocognitive, and dispositional resources to communicate different types of meaning (i.e., situational, sociolinguistic, sociocultural, psychological, literary, rhetorical, and interactional meanings) in a highly contextualized domain of language use. Therefore, when speaking performance is modeled, it may not always be clear where the line between task characteristics and contextual features is. This is why an expanded notion of task characteristics, which includes relevant contextual features, has been adopted in the current study.

As models of language proficiency have evolved over time, views of language ability as a trait have shifted to ones of language performance as a product of several interacting components, which can be roughly grouped into two categories: personal attributes and task characteristics. Figure 2.6 highlights these interactions in a simplified model of speaking performance. In the model, personal attributes include all the resources available to learners, including their language knowledge, their individual experiences, their cognitive capacities, and affective variables. Task characteristics here include not only characteristics of the task but also characteristics of the

context of the task. Accordingly, any variations in personal attributes or task characteristics could systematically influence speaking performance and would therefore be relevant to explaining the meaning of test scores (Bachman & Palmer, 1996).

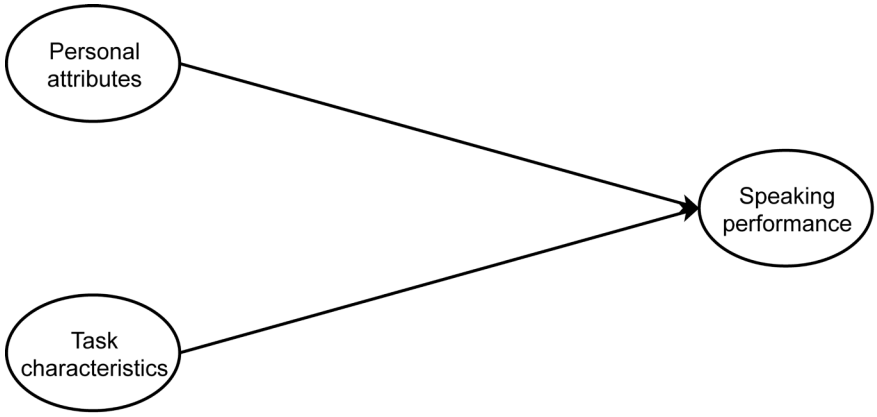


Figure 2.6. Simplified model of speaking performance.

How task characteristics affect speaking performance is an empirical question. Among adults, the difficulty of speaking tasks can be influenced by the topic or domain of the task (Khabbazbashi, 2017; Lumley & O'Sullivan, 2005), the speech functions elicited (Park, 2008; Weir & Wu, 2006), and even the presence or absence of visual stimuli (Elder, Iwashita, & McNamara, 2002; Iwashita, McNamara, & Elder, 2001). However, considering some of the age-related constraints of YLLs, such as their still-developing cognitive capacities, interactions with various multimedia task characteristics on a technology-enhanced speaking test are not only possible but quite likely. Furthermore, there is a need to see whether other cognitive processes, such as engagement, must be accounted for in models of performance. Some of these relationships are investigated in the current study.

The current study is situated in the context of building a partial validity argument for a test for YLLs. In particular, new evidence was gathered about the role of YLL age in evaluation, generalization, and explanation claims for a technology-enhanced speaking test. In the process, the effects of task characteristics hypothesized to support examinee engagement with the tasks

were examined both quantitatively and qualitatively. The following section summarizes some of the theoretical and empirical work that frames investigations into the nature of engagement with test content by YLLs.

Engagement

Like the simplified model of speaking performance (see Figure 2.6), engagement also results from interactions between certain personal attributes and task characteristics. After all, engagement describes "the level of involvement or connection between person and activity" (Ainley, 2012, p. 285). Figure 2.7 shows a synthesis between the simplified model of speaking performance and a model of engagement. In the model, personal attributes and characteristics of the task interact and spur engagement. One personal attribute that is relevant to engagement is motivation, which is the underlying psychological *trait* that turns into the *state* of engagement through "energized, directed, and sustained action" (Skinner et al., 2009, p. 225). Task characteristics that support engagement are a focus of the current study. The model also points to how engagement precedes and likely sustains speaking performance.

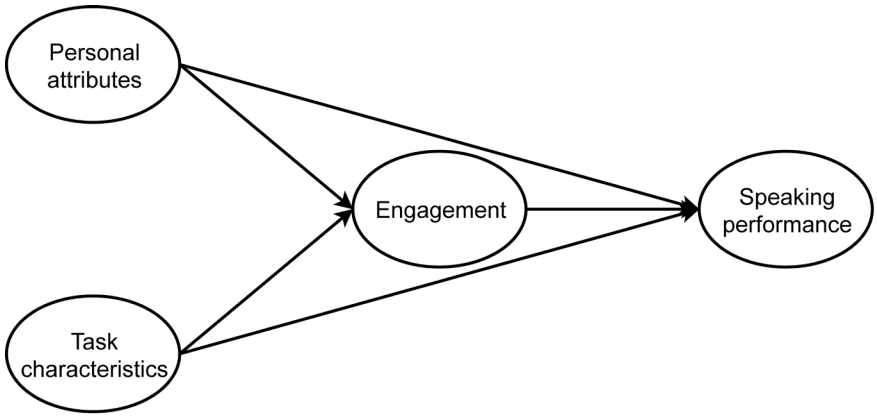


Figure 2.7. Simplified model of engagement and speaking performance.

Little can be done from a test development point of view to influence personal attributes, such as motivation, except, perhaps, trying to make sure that the consequences of engaging with a

test are positive and support motivations for further language use and learning. Practically speaking, however, in the context of developing a language assessment, only task characteristics can be manipulated. Task "engagement is malleable and is responsive to variations in the context" or characteristics of tasks (Fredricks & McColskey, 2012, p. 765), but questions remain about what task characteristics should be employed or avoided to make technology-enhanced speaking tasks more engaging for YLLs. Multimedia content has been widely cited as having the potential to enhance engagement as well as the power to distract (e.g., Gilakjani, 2012; Mayer, 2014).

Experience with educational games suggests that providing feedback about responses, offering incentives (e.g., game points and other rewards), ensuring appropriate task difficulty, supporting a sense of control (e.g., making choices), and creating interesting environments (e.g., aesthetics, multimedia, underlying characters, and storylines) can support engagement (McNamara et al., 2010). However, this list is divorced from any coherent theory of motivation or engagement and may well be incomplete. Therefore, a theoretically driven taxonomy of task characteristics that may support engagement is needed. In this section, after describing the theoretical underpinnings, such a taxonomy is proposed along with a brief review of approaches for measuring engagement.

Theories of Motivation and Related Models

To inform the development of a taxonomy of task characteristics that may support engagement, prominent theories of motivation and related models were reviewed for their connections to tasks. Since engagement and motivation are related constructs (Appleton et al., 2006), the use of the two terms in the literature significantly overlaps. Consequently, four theories or models in particular stood out as being clearly relevant to tasks and their characteristics: self-determination theory (Ryan & Deci, 2000, 2017), expectancy-value theory (Eccles, 1983; Eccles

& Wigfield, 2002), a model of interest development (Hidi & Renninger, 2006), and the ARCS model of instructional design (Keller, 1987, 2009). Each one is presented in turn.

Self-determination theory (Ryan & Deci, 2000, 2017) states that humans are driven by three fundamental needs. Tasks can be viewed as vehicles for meeting these needs. The first need is a need for competence. Competence involves demonstrating abilities to act capably and experience success in a variety of contexts. Second, there is a need for autonomy, which relates to agency, or having a sense of power and control (Skinner, 1996). Third, a need for connectedness involves relating socially to others. While task characteristics that support competence, autonomy, and relatedness may in turn support task engagement, this list seems incomplete.

Expectancy-value theory (Eccles, 1983; Eccles & Wigfield, 2002) is a complimentary theory of motivation that also directly relates to tasks. According to this theory, motivation is the product of one's expectancy for success and the value of the task. The notion of success here seems to overlap with the self-determination theory's need for competence. Task value is made up of four components: the attainment value, the perceived costs, the utility value, and interest. The attainment value reflects the intrinsic importance of doing well on the task. In contrast to the attainment value are the perceived costs, which relate to the time and effort needed for an activity as well as potential negative outcomes. Utility value describes the relevance of the task in the context of some greater short- or long-term goal, like winning a game or learning a language. And lastly, interest relates to the enjoyment of the activity. In terms of supporting a state of engagement, an expectancy for success, utility value, and interest seem most amenable to being influenced by characteristics of the task.

Another conceptualization of interest comes from Hidi and Renninger's (2006) model of interest development. In this model, there are two kinds of interest: personal interest and situational interest. Personal interest describes the kind of interest that is specific to individuals, such as topics and activities that individuals are motivated to engage in. Situational interest, on

the other hand, is not a personal attribute but rather stems from characteristics of the task. For example, situational interest can be created by novel and surprising task content. The model of interest development suggests that, over time, situational interest can turn into personal interest, but, for the purpose of developing test content, only situational interest is relevant.

Last is Keller's (Keller, 1987, 2009) ARCS model of instructional design. This model reflects an approach to facilitating motivation in the classroom. The acronym ARCS stands for attention, relevance, confidence, and satisfaction. Attention relates to capturing the interest of learners and stimulating their curiosity. Relevance is about meeting learners' needs and goals. Confidence involves supporting learner success and sense of control, and satisfaction is supported by reinforcing accomplishment with rewards. While by this point the reader will recognize that the ARCS model reiterates some themes suggested by the other models, it too seems incomplete.

Table 2.2 lists the components of the four theories and models presented side-by-side to show how components from the different theories and models align with each other. For example, the fundamental need for competence in self-determination theory roughly approximates the expectancy for success in expectancy-value theory, so they sit on the same horizontal plane. Looking across each row reveals that some categories, such as situational interest from the model of interest development, are clearly relevant to characteristics of the task. In this case, situational interest can be supported by task characteristics that spark interest in a particular situation. However, there are a few categories, such as attainment value and perceived costs from expectancy-value theory, that appear to be attributes of the learner and not characteristics of the task. By pulling out only those categories that reflect task characteristics that may support engagement, a taxonomy was revealed that informed the current study.

Table 2.2

Alignment of Motivation Theories and Relevant Models

Self-determination theory	Expectancy-value theory	Model of interest development	ARCS model of instructional design
Autonomy			Confidence
Competence	Expectancy for success		
Relatedness			
	Attainment value	Personal interest	Satisfaction
	Perceived costs		
	Utility value	Situational interest	Relevance
	Interest		Attention

A Proposed Taxonomy of Task Characteristics

The motivation literature suggests four kinds of task characteristics that may support engagement: those that support a sense of agency, those that support success, those that support social connectedness, and those that support situational interest. The categories of agency, success, and social connectedness are derived from the fundamental needs for autonomy, competence, and relatedness, respectively, from the self-determination theory. Success and situational interest correspond to the expectancy for success and two components of task value (i.e., utility value and interest) from expectancy-value theory. Situational interest also mirrors that from the model of interest development and the components of relevance and attention from the ARCS model of instructional design.

To help drive a research agenda and perhaps someday inform test development efforts, these categories were arranged into a taxonomy (see Table 2.3). Although most of the examples

can be treated as characteristics of the input, they can also all be intentionally manipulated to support engagement (Jones, 2009). Each of the four categories from the proposed taxonomy is discussed here in turn.

Table 2.3

A Taxonomy of Task Characteristics That May Support Engagement

Engagement			
Task characteristics that support a sense of **agency**	Task characteristics that support a sense of **success**	Task characteristics that support **social connectedness**	Task characteristics that support **situational interest**
e.g., topical choice, task choice, empowering role-play	e.g., vocabulary support, task scaffolding, earning points, appropriate challenge	e.g., teamwork, interaction, competition, shared experiences	e.g., novelty, video animation, avatars, storylines, games, humor, relevance to goals

Agency. Task characteristics that facilitate a sense of agency, of power and control (Skinner, 1996), are expected to support engagement among YLLs. This can be accomplished by providing choices about and within tasks, including topical choice. The beneficial effects of choice on young learner engagement have long interested researchers (e.g., Cordova & Lepper, 1996), though too many choices may also inhibit rather than enhance engagement (Iyengar & Lepper, 2000). Unfortunately, studies on the effects of topical choice on language test tasks or more specifically on speaking tasks are few in number.

One study about topical choice that did involve a language test was conducted by Jennings, Fox, Graves, and Shohamy (1999). The Canadian Academic English Language Assessment was administered to 254 university applicants under two different conditions. Half the test takers were allowed to choose from several versions of the test that focused on different topics. The other half was assigned one of the versions of the test with no choice about the topic.

Although speaking tasks were not part of the test, results indicated no significant performance advantages for either condition. Even though 72% of the test takers reported a preference for having a choice of topics when taking a test, it is not clear whether having choice actually resulted in greater engagement, which was not specifically measured.

Outside of the testing environment, Mozgalina (2015) reported on an investigation of 72 adult learners of Russian at a German University who needed to work in pairs to prepare a presentation about a famous person from Russia. The effects of manipulating choices regarding the task content were examined by giving learners no choice, limited choice, or free choice in terms of the topic (person) for their presentations. Post-task questionnaires revealed that choice in terms of task content seemed to heighten engagement. However, in a second investigation with 24 pairs of learners, the effects of giving procedural choices (i.e., by not obligating the use of a set of guiding questions) were evaluated. Results suggested that the addition of procedural choice actually detracted from participant engagement with the task.

Thurman (2013) also investigated university students learning a second language and the effects of choice on them. The 143 language learners were exposed to an oral information gap task under one of two conditions: having no choice or being able to choose one of three topics. A survey of participants afterward indicated that there was increased interest in the task when performing it under the choice condition. A follow-up study of a smaller subset of 37 pairs also revealed that the complexity of the language produced as well as time on task was greater for the choice condition.

Lambert, Philp, and Nakamura (2017) also investigated how choice affects performance and engagement by administering picture narration tasks to 32 English majors at a Japanese university. Students used four-picture sequences to tell stories under two conditions: in the no-choice condition, the pictures were assigned by the teacher, and in the choice condition, the pictures were generated by the students. The spoken responses were recorded, and the time on

task as well as the number of words, elaborative clauses, negotiations moves, and backchannels produced were determined. Each of these measures was greater under the choice condition regardless of ability level. A follow-up questionnaire with 12 Likert-type items about participants' experiences with the tasks also suggested that the choice condition resulted in greater engagement.

Although choices, even superficial ones, have long been recognized as supporting learner engagement (e.g., Cordova & Lepper, 1996), choices may not always benefit performance (Iyengar & Lepper, 2000; Mozgalina, 2015). Since most studies involving language learners and choice have focused on adult learners, how generalizable findings are to the context of YLLs is not clear. Therefore, the effects of topical choice on YLLs' engagement and speaking performance were investigated in the current study.

Success. "When young learners are assessed, it is important that children experience overall success" (McKay, 2006, p. 14). One speaking task characteristic hypothesized to support a sense of success among YLLs is providing vocabulary support, such as explicitly providing vocabulary relevant to completing a task. By seeing pictures, reading text, and hearing the pronunciation of select lexical items that might be useful to completing a task, examinees are enabled to show what they can actually *do* with the language instead of whether they know a few key words. The words presented could be optional (in that their use would not be required in order to complete or receive the highest scores on the task).

Vocabulary support is one example of how scaffolding techniques can be adapted in the context of a large-scale assessment (Hauck, Pooler, Wolf, Lopez, & Anderson, 2017). Such approaches can support a sense of success by assisting "a child or novice to solve a problem, carry out a task or achieve a goal which would be beyond his [or her] unassisted efforts" (Wood, Bruner, & Ross, 1976, p. 90). Vocabulary support in the form of multimedia glosses, or presenting the meaning of unknown words with text, pictures, or video, has been shown to help

adults with reading and listening comprehension tasks (e.g., Jones & Plass, 2002; Yanguas, 2009). However, the effects of vocabulary support on supporting a sense of success or on engagement with speaking tasks among YLLs were not clear.

However, in a series of studies that incorporated eye-tracking technologies, Ballard and Lee (2015) and Lee and Winke (2017) investigated how YLLs and native speakers interact with *TOEFL Primary* Speaking test tasks, including tasks that provide vocabulary support. Several computer-delivered speaking tasks were administered to 24 and 28 children, respectively. The researchers noticed that YLLs spent more time than native speakers did attending to vocabulary words presented as labeled pictures on the screen, suggesting a greater level of engagement with that task characteristic among YLLs. They also observed that the amount of language generated (number of syllables), time on task, and articulation rates were greater for both groups on a task that introduced relevant vocabulary compared to a task that lacked vocabulary support. However, other major differences between the two tasks, such as the speech function elicited and whether a video animation was present, could also account for these differences. Unfortunately, the studies do not report what young learners may have revealed during post-hoc interviews about the tasks or other task characteristics. The current study further explored how vocabulary support on these *TOEFL Primary* Speaking tasks would affect YLLs in terms of engagement and performance.

Social connectedness. Relating to others is particularly important to young learner engagement (Furrer & Skinner, 2003). Relating to others can involve working together to negotiate a task, competing in a task, or even just having a shared experience. In a language assessment context involving relating to an interlocutor, the interlocutor is frequently the interviewer or, less commonly, a peer. Studies have suggested that interlocutor variation can have a significant effect on performance among adults (e.g., Brooks, 2009; Davis, 2009; Galaczi, 2008) and with YLLs (Kondo-Brown, 2004). The effects of interlocutors or the dynamics of interactions with them on engagement are less clear.

In terms of technology-enhanced assessments, there are also possibilities for interaction. It should come as no surprise that video-game players, especially those who play multiplayer online games, report feeling like they are relating to other players. What is surprising is that when games include non-person characters—virtual characters who may be even more responsive than a player's real-life peers—there can be an even greater sense of feeling like one is relating to others (Rigby & Ryan, 2011). Virtual characters, or avatars, can also be used in a technology-enhanced assessment to provide authentic reasons to engage in communication (Cho et al., 2016, 2017). The effects of including such features on a speaking test on performance and engagement go beyond the scope of the current study, but they would make interesting research topics for another time.

Situational interest. Task characteristics can support engagement by stimulating situational interest. Unlike personal interest, which can vary considerably between individuals, situational interest can be sparked by the characteristics of the situation or task (Krapp, Hidi, & Renninger, 2014). For example, situational interest can be created by novel or surprising task content. Another way to generate interest in a task is by explicitly connecting it to learners' lives and goals (Hulleman, Godes, Hendricks, & Harackiewicz, 2010; Vansteenkiste, Lens, & Deci, 2006).

Two task characteristics hypothesized to support situational interest in the context of a speaking test for YLLs are novelty and video animation. Novel situations, which tend to be playful and would not occur in real life, have been identified by YLLs as a feature of educational games that would stimulate interest (Butler, 2017b). The use of video animation (in contrast to static images) has also been shown to have a positive effect on efforts by very young language learners (Verhallen & Bus, 2009). Although novelty and video animation have long been believed to stimulate interest in students (Bergin, 1999; Malone, 1981), their effects on YLL engagement

and speaking performance are less clear. The relationships among these two task characteristics, age, engagement, and performance were a focus of the current study.

Measures of Engagement

In order to evaluate the influences of various task characteristics on speaking task engagement, reliable indicators of engagement are needed. Engagement appears to be a multidimensional construct (Fredricks, Blumenfeld, & Paris, 2004; Svalberg, 2009), though there are varying opinions about how many dimensions there are and how to distinguish among them (Philp & Duchesne, 2016; Reschly & Christenson, 2012). Two commonly cited dimensions that are particularly relevant at the task level are emotional (or affective) engagement and cognitive engagement. Emotional engagement involves a level of excitement and arousal that may relate to attitudes about the task, such as liking or having an aversion toward it. Cognitive engagement relates to attention and the amount of mental effort devoted to the task.

Identifying indicators of emotional and cognitive dimensions of engagement has been a challenge for researchers, so many different approaches to measuring engagement have been employed (Fredricks & McColskey, 2012; Henrie, Halverson, & Graham, 2015). For example, self-report surveys are routinely used to gauge affective and cognitive engagement because they are more direct than teacher surveys or observations (Appleton et al., 2006). The usefulness of surveys of YLLs, however, may be limited. Age-related metacognitive constraints (Lai, 2011) and limitations in language (Hasselgreen & Caudwell, 2016) could prohibit some YLLs from being able to articulate the nature of relationships between task characteristics and various states of engagement. Verbal reports such as those obtained through retrospective interviews can also be employed to probe YLLs about their engagement with tasks and task characteristics (see Ericsson & Simon, 1980, 1993). However, verbal reports tend to be resource intensive and time consuming. Therefore, complementary indicators of task engagement should also be considered to help triangulate findings (Greene, 2015).

There are several approaches for measuring engagement beyond self-reports. Sometimes, subjective teacher or parent reports may be able to provide useful information about young learner engagement (e.g., Oga-Baldwin, Nakata, Parker, & Ryan, 2017), but these reports alone may not be very reliable. Objective physiological measures, such as electroencephalograms (EEGs) and skin conductance tests, can also be considered as potential measures of engagement (D'Mello, Dieterle, & Duckworth, 2017). However, these approaches can be costly and tend to be obtrusive for YLLs. Yet another approach is using eye-tracking technologies, which involve analyzing participant gaze to infer what the participant is attending to (or engaged with) on a screen (Winke, 2015). This approach, though, can also be costly, and its usefulness would be limited when evaluating potential engagement with nonvisual stimuli. Nonverbal behaviors and facial expressions, such as leaning forward, smiling, and blinking, have also been considered as indicators of engagement (e.g., Castellano et al., 2010; D'Mello & Graesser, 2010; Hsieh, Lin, & Hou, 2014).

Other alternatives that are particularly relevant to speaking tests involve examining qualities of spoken responses themselves for clues about task engagement. For example, time on task and the amount of language generated have been used as proxies for cognitive engagement in the context of speaking tasks (e.g., Dörnyei & Kormos, 2000; Lambert et al., 2017; Phung, 2017). More time spent engaged with a task suggests greater cognitive involvement with it, although the depth and intensity of that involvement may not be accounted for. In addition, the amount of language generated may conflate levels of cognitive engagement and language ability.

Other potential behavioral indicators of engagement that are unique to spoken language involve paralinguistic acoustic features. These acoustic features, including acoustic indicators of emotion (e.g., Banse & Scherer, 1996; Eyben, 2016; Scherer, 2003), measures of fluency (Kormos, 2006), and features of prosody (Kalathottukaren, Purdy, & Ballard, 2015), regularly accompany speech. For example, measures related to pitch (or its acoustic equivalent, F_0 or

fundamental frequency) and speech rate have routinely been identified as indicators of emotion and arousal (e.g., Banse & Scherer, 1996; Bänziger, Hosoya, & Scherer, 2015; Juslin & Laukka, 2001; Laukka et al., 2016; Truesdale & Pell, 2018). Typically, machine learning algorithms, such as support vector machines, are used to make classification decisions based on a set of acoustic features (Eyben, 2016).

In terms of using acoustic features to signal engagement, only a few small-scale studies have been conducted involving engagement in conversation and computer interactions. For instance, Yu, Aoki, and Woodruff (2004) explored the use of measurements related to F_0, energy (i.e., loudness), the duration of voiced segments, and formants (which distinguish vowel sounds from one another) to evaluate engagement in everyday adult conversations using samples from two corpora: The Linguistic Data Consortium's Emotional Prosody corpus, which is composed of speech samples by voice actors emulating different emotions, and the CallFriend corpus, which consists of samples of social telephone calls that were then coded for emotive content. The researchers found that classification accuracy was better for the acted Emotional Prosody corpus samples than the spontaneous samples in the CallFriend corpus, though it is not very clear what emotional states (or levels of arousal and valence) were treated as signs of engagement. Another finding was that acoustic measures that were most useful for detecting arousal varied between genders, suggesting that, in terms of detecting engagement among YLLs, there may be a need to account for variations between speakers.

Variations between groups in terms of the acoustic features that were most useful for detecting engagement were also observed in a study by Gupta, Bone, Lee, and Narayanan (2016). During the course of the development of the Rapid ABC screener for autism, they investigated acoustic measures related to F_0, energy, jitter (moment-to-moment perturbations in F_0), and shimmer (moment-to-moment perturbations in energy) in recordings of 63 children under 3 years of age as they interacted with four adult psychologists. Behavior and engagement levels were

annotated by the psychologists. Acoustic measures that were most useful for detecting interactional engagement varied between the children and their interlocutors (i.e., the psychologists), and the cues from interlocutors were actually more useful for detecting engagement. This suggests that variations in age may also play a role in terms of the sensitivity of different acoustic measures as signals for engagement.

In a comparison of child-to-child and child-to-robot engagement among 62 children between 5 and 10 years old, Chaspari and Lehman (2016) examined measures related to F_0 and energy of certain keywords that were uttered in the course of playing a computer game. Audiovisual recordings of each child were then reviewed in ten-second intervals for the presence of engagement. Analyses revealed that when children played the game with other children instead of with the robot, which did not vary the pitch and energy in its responses to the children's utterances, the acoustic cues were stronger predictors of engagement. This suggests that engagement, like language, may be co-constructed and that the level of engagement or its influence on acoustic cues may be muted in interactions with nonhuman agents. This may have implications for engagement detection on a technology-enhanced speaking test for YLLs.

Even in human-to-human interactions, there can be variations in the quality of the interaction that affects the sensitivity of acoustic features to engagement. Kim and Truong (2016) reported on a study of 21 Dutch children between 5 and 8 years of age who worked in groups of three to collaboratively build a 3D puzzle. Videos of the activity were coded for signs of harmonized (i.e., more collaborative) and unharmonized (i.e., less collaborative) engagement, and speech was analyzed in terms of turn behavior, F_0, energy, zero-crossing rate (which commonly distinguishes voiced from unvoiced speech), harmonicity (or the logarithmic harmonic-to-noise ratio, which is a measure of the intensity of harmonic waveforms), jitter, and shimmer. Results indicated that the energy, harmonicity, and zero-crossing rate showed more variability during unharmonized engagement, and energy was greater. This further suggests that a social element is

relevant to engagement and its detection with acoustic features. As virtual agents become more sophisticated (i.e., more responsive), the potential for using acoustic features as a signal for engagement may increase, including on technology-enhanced assessments for YLLs.

Research into using acoustic features as signals of engagement is limited, and so is the number of acoustic features that have been evaluated as potential signals. The current study expanded upon the range of acoustic features considered as potential indicators of engagement (see Appendix D for a list of 100 features considered). With dependable indicators of engagement, the influence of task characteristics hypothesized to support engagement could become more clear, and the resulting effects on speaking performance on a test for YLLs could be better understood.

Despite calls for engaging assessment tasks (e.g., Hasselgreen, 2000, 2005; Hauck et al., 2013; McKay, 2006; Taylor & Saville, 2002; Wolf & Butler, 2017), in practice, the development of tasks with engaging characteristics has depended heavily on expert judgment rather than empirical data. The current study sought to begin changing that by first proposing a theoretically driven taxonomy of task characteristics that may support engagement and then exploring ways to identify the effects of those task characteristics on engagement and the effects of engagement on speaking performance in the context of a technology-enhanced test for YLLs.

Summary

This literature review highlighted some gaps in the research regarding age-related issues that may be involved in the development of YLL assessments, the validity of speaking assessments for YLLs, and factors and indicators of engagement. For example, it was not clear how examinee age might interact with tasks and raters. The effects of task characteristics such as topical choice, vocabulary support, novelty, and video animation on measurement quality have not been well studied. And the use of paralinguistic acoustic features to better understand the

nature of engagement needed more exploration. The goal of this study was to examine these topics in the context of tasks developed for Educational Testing Service's *TOEFL Primary* Speaking test.

Chapter III

METHODOLOGY

This chapter describes the methods employed to address the research questions. The study followed a mixed-methods approach, with a quantitative phase and a qualitative phase. The quantitative phase involved gathering evidence related to the evaluation, generalization, and explanation inferences of a validity argument (see Chapelle et al., 2008, 2010; Kane, 1992, 2006, 2013; Knoch & Chapelle, 2018) for a technology-enhanced speaking test for YLLs. More specifically, evidence was gathered about task and rater functioning across age groups, about the optimal configuration of tasks and ratings to maximize score dependability, and about how task characteristics hypothesized to support engagement influenced performance. Findings about the nature of the relationships between these task characteristics and engagement were then triangulated in the qualitative phase.

For the quantitative phase, a non-experimental, ex post facto research design (Grotjahn, 1987) was used. Response data from a pilot administration of Educational Testing Service's *TOEFL Primary* Speaking test were obtained, and the analyses were divided into four stages corresponding to the first four research questions they addressed. The purpose of Stage 1 was to investigate how well the test (i.e., its tasks, raters, and rating scales) discriminated among YLLs of varying speaking abilities, including whether there were any systematic interactions between tasks and examinee age groups or raters and examinee age groups. In Stage 2, the degree to which response scores were generalizable between parallel tasks and ratings and, by extension, between different forms or versions of the same test were explored. Stage 3 revealed how task characteristics hypothesized to support engagement (i.e., topical choice, vocabulary support, novelty, and video animation) affected measurement qualities like task difficulty. The last stage

sought to disentangle the effects of task characteristics on engagement from their effects on performance.

The qualitative phase of the study followed a qualitative, multi-case study research design (Wiersma & Jurs, 2008). Videotaped interview data were gathered from a small sample of YLLs in order to address the last research question. The purpose of the qualitative phase was to substantiate the findings of Stage 4 from the quantitative phase. More specifically, feedback from YLLs about the tasks and their characteristics were needed to confirm the effects of topical choice, vocabulary support, novelty, and video animation on engagement as indicated by paralinguistic acoustic features.

The Quantitative Phase

Participants

Over 400 YLLs from around the world took part in the pilot administration of *TOEFL Primary* Speaking test tasks in December 2012. Since no official scores were to be generated, the test was framed as a low-stakes activity for students. Several participants were eventually excluded from the study because of reported or suspected technical difficulties, such as malfunctioning microphones, leaving 401 participants. The remaining participants hailed from 11 different countries (Brazil, China, Egypt, Japan, Jordan, Lebanon, Morocco, Peru, South Korea, Tunisia, and Vietnam), and approximately 53.4% of them reported being female. Participants ranged from 7 to 13 years of age, with a mean age of 10 years 4 months, a median age of 10 years 3 months, and a standard deviation of about 16 months. The participants were classified as belonging to one of three approximately equal groups according to age (there were 131 and 127 participants in the youngest and oldest groups, respectively) for later between-group analyses.

Instruments

A pre-test demographic questionnaire, the *TOEFL Primary* Speaking pilot test, and a post-test engagement survey were used in the current study.

Pre-test demographic questionnaire. The purpose of the pre-test demographic questionnaire was to gather background information about the examinees. The brief questionnaire was administered on a computer immediately before the test was administered. It asked for basic demographic information about each participant and for details about each participant's English-learning experience. In the current study, only the information related to participants' ages was used. The pre-test demographic questionnaire is presented in Appendix A.

The *TOEFL Primary* Speaking pilot test. The *TOEFL Primary* Speaking test is intended to measure young English-as-a-foreign-language learners' abilities to communicate orally in routine social situations related to their daily lives (Cho et al., 2016, 2017). It is computer-delivered and semi-direct, which means that the responses are recorded and later scored by human raters (Clark, 1979). The tasks are designed to elicit performances that provide evidence of speaking ability, and the construct is operationalized through tasks that obligate the use of six targeted speech functions, or "communication goals" (Cho et al., 2016, 2017). The speech functions, which are commonly found in curricula around the world (Turkan & Adler, 2011), are listed in Table 3.1[1]. Although operational forms consist of seven scored tasks, the pilot contained eleven scored tasks testing the six speech functions. Each task was built around the scenario of visiting a zoo. Descriptions and screenshots of the tasks are presented in Appendix B.

[1] Although Educational Testing Service distinguishes "asking questions" and "making simple requests" as distinct communication goals tested by the *TOEFL Primary* Speaking test, the associated tasks functioned so similarly that the two categories have been collapsed into one for the purpose of the current study.

Table 3.1

Correspondence Between Elicited Speech Functions and Task Types

Speech Function (Cho et al., 2016, 2017)	Task Type
Giving directions	Give Directions
Explaining and sequencing simple events	Retell a Story
Asking questions/making simple requests	Ask Questions
Giving simple descriptions	Describe a Picture
Expressing basic emotions, feelings, and opinions	Express an Opinion

Post-test engagement survey. The purpose of the post-test engagement survey was to gather some feedback about the test from YLLs themselves. A simple survey with two dichotomous items was presented on the computer immediately after the test was administered. The items asked examinees about their engagement with the test tasks. Examinees were asked to either agree or disagree with each of the following statements:

1. I enjoyed taking this test.
2. The zoo story was fun.

Procedures

Test administration procedures, rubric development processes, and scoring procedures are described below. In addition, how tasks were coded for task characteristics and how acoustic features were extracted and selected as indicators of engagement are also described.

Test administration. The assessment was administered to individual students using custom-built software on computers with headphones and microphones. The test was self-contained, meaning that once it began, it ran all the way through with no intervention needed from the test taker or a supervising adult. There were no opportunities to redo a response or skip ahead to future prompts. After each prompt, a countdown timer on the screen indicated how much recording time remained, and a beep indicated when to stop speaking. Preparation time was not provided in order to better emulate real-life conversation, and the response time varied according

to the task. The first task was an unscored warm-up task. All responses were recorded as 16-bit WAV files at 44.100 kHz for later distribution and scoring.

Rubric development. Holistic scoring guides, or rubrics, describing the rating scales were developed and refined through the consensus of a small team of test development experts who listened to a random subset of responses for each task type, grouped them into natural categories according to performance level, and then described the characteristics of each group of responses in terms of their overall achievement of the targeted speech functions, including features related to meaningfulness, completeness, and intelligibility. These descriptors reflect a bottom-up approach to rubric development in which the salient and distinguishing features of different groups of responses were described instead of a top-down approach in which the rubric descriptors were derived from a specific model of performance or proficiency. However, it is not difficult to see how the descriptors relate to language resources, including grammatical, propositional, and functional knowledge, specified in recent models (e.g., Purpura, 2017). The varied length and complexity of responses to tasks testing different speech functions warranted the creation of two sets of rating scales. For extended responses to *Retell a Story* or *Give Directions* tasks, a 0-to-5-point rating scale was developed. For responses to the other tasks, a 0-to-3-point rating scale was developed. Scoring guides describing both rating scales can be found in Appendix C. In addition to the rubrics, a few benchmark samples exemplifying each score point for each task type were selected.

Response scoring. Responses were scored by a team of nine experienced raters who were instructed to carefully review the rubrics and the benchmarks at the start of each scoring session and then again as often as needed. Each response received two blind ratings, meaning that the second rater did not see the previously assigned score. All scores were recorded electronically.

Task coding. The 11 test tasks were independently coded for the presence or absence of topical choice, vocabulary support, novelty, and video animation by two test development experts. There was a 100% agreement rate between coders, suggesting that the task characteristics were quite salient. Table 3.2 shows how tasks were coded for each task characteristic.

Table 3.2

TOEFL Primary *Speaking Tasks and Their Characteristics*

Task	Task Type	Topical Choice	Vocabulary Support	Novelty	Video Animation
1	Express an Opinion	X	X		
2	Give Directions				
3	Describe a Picture		X	X	
4	Describe a Picture			X	
5	Retell a Story		X	X	X
6	Ask Questions				
7	Ask Questions	X			
8	Retell a Story		X	X	
9	Ask Questions				
10	Give Directions				
11	Express an Opinion	X			

Note. An "X" indicates that the task characteristic was present.

Extracting and selecting acoustic features. Identifying likely indicators of engagement involved many steps. Two sources of data were available: the results of the post-test engagement survey and audio files of the spoken responses themselves. Basically, the strategy involved examining relationships between the results of the post-test engagement survey and acoustic (and prosodic) features extracted from the responses. Features that pointed to between-subject differences in engagement at the level of the whole test were used to infer within-subject differences in engagement at the level of individual tasks. Each step of the process, including feature extraction and two consecutive rounds of feature selection, is described in detail below.

Audio file data needed to first be prepared for feature extraction. In order to focus on the portions of the audio files in which examinees were engaged with the tasks, any silent periods

after spoken responses had ended were automatically trimmed from the ends of each of the 4,411 audio files using Praat (Version 6.04; Boersma & Weenink, 2016). Then, the remaining audio files for each examinee's responses were concatenated into a single longer audio file using SOund eXchange (Version 14.4.2; Bagwell, 2015). In the end, there were 12 audio files for each participant—one for each response to the 11 scored tasks on the instrument and one longer one containing all 11 responses concatenated together.

From each of the 4,812 audio files, 100 acoustic and prosodic features were extracted as possible indicators of engagement. Appendix D contains the complete list of features extracted. These features were measures related to pitch, timbre, energy, duration, and tempo, among others, and they included features that had previously been identified in the literature as being related to engagement (see Chapter 2). The features were extracted using two programs. OpenSmile (Version 2.30; Eyben, Weninger, Gross, & Schuller, 2013) was used to extract the eGemaps feature set, which contains 88 features widely believed to reflect emotional states that may be encoded in speech (Eyben et al., 2016). Praat (Version 6.04; Boersma & Weenink, 2016) was used to extract 12 additional features also hypothesized as possible indicators of engagement. In all, almost half a million measurements were taken.

To narrow down the list of 100 possible indicators of engagement to a shorter list of potential indicators, correlations between the sum of positive responses to the post-test engagement survey and the features extracted at the whole-test level (i.e., from the concatenated audio files) were evaluated. Since some features were likely influenced by age and gender (e.g., features related to pitch) and by ability (e.g., speech rate), linear regression was performed using SPSS (Version 23; IBM Corp., 2015) for each feature with age, gender, ability estimates, and engagement survey scores as independent variables. After controlling for age, gender, and

ability, only nine features had significant correlational relationships with reported test engagement[2].

The nine potential indicators of engagement identified during initial feature selection were further narrowed down to identify which were likely to be the most robust. To do this, the measures from individual responses were transformed into *z* scores for each individual examinee using the "split file" and "standardize" commands in SPSS (Version 23.0; IBM Corp., 2015). This procedure muted the influences of examinee attributes like age, gender, and ability, which presumably would not change between tasks during the same test administration, thus making relative differences between responses from each examinee more salient. Then, the nine transformed features were modeled as indicators of a single latent factor (i.e., task engagement) using Mplus (Version 8; Muthén & Muthén, 2017). The R^2 values of these potential indicators were examined, and three indicators with the highest R^2 values were retained as the most likely indicators of engagement for subsequent model building. The final indicators were the mean and standard deviation of shimmer (i.e., moment-to-moment perturbations in energy) and the mean harmonicity (i.e., the logarithmic harmonic-to-noise ratio). Perceptually, shimmer relates to the breathiness of speech, and low harmonicity is commonly associated with hoarseness (Teixeira & Fernandes, 2014). These features had previously been identified in the literature as potential indicators of engagement among children (Gupta et al., 2016; J. Kim & Truong, 2016).

Data Analysis

The analyses of the data were divided into four stages.

---

[2] These features were the mean syllables per run, the 20th percentile of F_0, the mean and standard deviation of shimmer, the mean harmonicity, the standard deviation of the harmonic difference H1-H2, the mean frequency of F_1, and the standard deviations of the frequencies of F_2 and F_3. See Appendix D for more details about the meanings of these features.

Stage 1: Investigating the evaluation of spoken responses. The first stage of the analyses addressed the first research question about how well the test discriminates among YLLs according to their speaking abilities. Evidence related to how scores to responses reflect targeted language abilities was gathered in the context of the evaluation inference in a validity argument. This stage involved fitting response scores to a many-facet Rasch measurement model. Rasch analyses involve modeling probabilities of measurements as logistical functions of multiple parameters, or facets (Bond & Fox, 2015). Unlike other approaches to item response theory (e.g., Birnbaum, 1968), Rasch models are confirmatory and prescriptive (Bond & Fox, 2015), meaning that the data are expected to fit the theoretical model. Model fit indicates how well tasks, raters, and rating scales functioned.

A grouped rating scale model (Andrich, 1978) was specified with three principal facets—examinee ability, task difficulty, and rater severity:

$$\ln (P_{ngijk} / P_{ngijk\text{-}1}) = B_n - C_j - D_{gi} - F_{gk}$$

where

P_{ngijk} = the probability of examinee n, when rated on task i by rater j, being awarded a rating of k
$P_{ngijk\text{-}1}$ = the probability of examinee n, when rated on task i by rater j, being awarded a rating of k-1
B_n = the ability of examinee n
C_j = the severity of rater j
D_{gi} = the difficulty of task i in rubric group g
F_{gk} = the difficulty of achieving a score in category k relative to category k-1 in rubric group g

The analyses were performed using Facets (Version 3.22; Linacre, 1999). The convergence criteria for joint maximum likelihood estimation were set at no marginal score point residual greater than 0.5 score points and no log-odds, or logit, estimate changing faster than 0.005 logits per iteration. Examinee ability, task difficulty, and rater severity were mapped onto an equal-interval logit scale, and the resulting statistics were interpreted to describe how well the tasks, raters, and rating scales performed and were able to discriminate among examinees of varying speaking abilities.

Differential facet functioning (DFF) was also used to examine the possibility of task and rater bias in relation to age. “DFF refers to the simple observation that an item, a topic, a rater, or

other testing facet displays different statistical properties in different group settings" (Du, Wright, & Brown, 1996, p. 1). In order to conduct DFF, the original Rasch measurement model was expanded to include a dummy facet to indicate examinee membership in the youngest and oldest age groups. The dummy facet was anchored at zero so that it would be excluded as part of the measurement model, but it could still be used to explore potential interactions with other facets like task difficulty and rater severity. This approach addressed whether tasks functioned differentially across age groups and also whether raters were more lenient toward younger learners, perhaps because of the "cuteness" of younger examinees' responses.

Stage 2: Investigating the generalizability of scores. Following the many-facet Rasch measurement analyses, a series of generalizability and decision studies (Brennan, 2001a; Cronbach et al., 1972; Shavelson & Webb, 1991) was conducted to determine the components of score variance attributable to examinees, ratings, and tasks within each task type, the effects of varying the configuration of tasks and ratings on overall score dependability, and whether the relative contributions of task types to the composite universe score variance was similar across age groups. Stage 2 resulted in evidence to support assumptions about how consistent scores are over parallel versions of tasks and test forms, across ratings, and among examinees of different ages in the context of a generalization inference in a validity argument.

Using mGENOVA (Version 2.1; Brennan, 2001b), multivariate generalizability and decisions studies were conducted to address the second research question about the generalizability of scores. In the analyses, the object of measurement was persons (p). Tasks (t) and ratings (r') were considered random facets because they are merely samples from a universe of admissible tasks and ratings. Note that, consistent with the work of Lee (2005, 2006) and Schmidgall (2017), ratings (r') were used here instead of the more traditional raters (r) because the rating design was not fully crossed. The fixed facet was task type, which aligns with the speech functions targeted by the tasks (see Table 3.1). Persons and ratings were crossed with this

fixed facet, and the tasks were nested within the five task types. The design for the generalizability study, using Brennan's (2001a) notation in which a filled circle indicates crossing and an open circle indicates nesting, was as follows:

$$p\bullet \times t\circ \times r'\bullet$$

The generalizability study revealed the relative contributions of examinees (persons), tasks, ratings, and the myriad interactions between them to total score variability. However, generalizability theory can also be used to "evaluate the effectiveness of alternative designs for minimizing error and maximizing reliability...in a manner analogous to the Spearman-Brown prophecy formula in classical test theory" (Shavelson & Webb, 1991, p. 12). Decision studies predicted the effect of varying the number of ratings and the numbers of tasks of each task type on the dependability of scores. The pursuit of higher score dependability, as indicated by the phi (ϕ) coefficient, must be balanced with practical needs to keep rating costs down and to constrain test length to accommodate the limited cognitive and attentional resources of young examinees. Finally, the relative contributions of tasks and task types to the composite universe score variance were compared across age groups to evaluate the appropriateness of the configuration across the spectrum of examinee ages targeted by the test.

Stage 3: Investigating an explanation into the meaning of scores. The influences of certain task characteristics on measurement qualities like task difficulty were then investigated in Stage 3. To accomplish this, Fischer's (1973, 1995) linear logistic test model (LLTM) was employed. LLTM is an expansion of the Rasch model in which a single task facet is replaced by multiple facets for tasks' component characteristics. Therefore, instead of a task difficulty measure, a difficulty measure attributable to each task characteristic can be determined. Such an approach had previously been applied in the context of other language tests (e.g., Baghaei & Ravand, 2015; Gorin, 2005; Rahman & Mislevy, 2017; Sonnleitner, 2008). Stage 3 provided evidence to support assumptions about the meaning of scores in the context of an explanation

inference in a validity argument. More specifically, the results are relevant to current theoretical models of speaking ability, which postulate the systematic influence of task characteristics on performance.

LLTM built upon the Rasch analyses conducted in Stage 1 by decomposing the effects of task characteristics on performance. The task characteristics under scrutiny were the task types (corresponding to the speech functions elicited) and presence or absence of four task characteristics hypothesized to support engagement: topical choice, vocabulary support, novelty, and video animation. Again, using Facets (Version 3.22; Linacre, 1999), a grouped rating scale model (Andrich, 1978) was specified as follows:

$$\ln (P_{ngijk} / P_{ngijk\text{-}1}) = B_n - C_j - Q_{gi} - R_i - S_i - T_i - U_i - F_{gk}$$

where

P_{ngijk} = the probability of examinee n, when rated on task i by rater j, being awarded a rating of k
$P_{ngijk\text{-}1}$ = the probability of examinee n, when rated on task i by rater j, being awarded a rating of k-1
B_n = the ability of examinee n
C_j = the severity of rater j
Q_{gi} = the difficulty of the speech function in task i in rubric group g
R_i = the difficulty of the presence of **topical choice** in task i
S_i = the difficulty of the presence of **vocabulary support** in task i
T_i = the difficulty of the presence of **novelty** in task i
U_i = the difficulty of the presence of **video animation** in task i
F_{gk} = the difficulty of achieving a score in category k relative to category k-1 in rubric group g

The convergence criteria for joint maximum likelihood estimation were set at no marginal score point residual greater than 0.5 score points and no logit estimate changing faster than 0.005 logits per iteration. Examinee ability estimates, difficulty estimates for each task characteristic, and rater severity estimates were mapped onto a logit scale and the resulting statistics interpreted. The influences of task type, topical choice, vocabulary support, novelty, and video animation on task performance became clear. Measures of examinee ability, task type difficulty, and rater severity were retained for Stage 4 of the quantitative analyses.

Measurement qualities in addition to task difficulty were also determined. The discrimination index is a value that indicates how well a task discriminates between masters and non-masters. In item response theory, this value is commonly thought of as the slope of the item

characteristic curve. Since an assumption of Rasch measurement is that discrimination is consistent between items, discrimination was calculated by dividing the difference between the average task scores for the highest and lowest performing thirds of examinees by the number of score points available for the task. Closely related to discrimination indices are point-measure correlations (Bond & Fox, 2015), which are similar to point-biserial correlations, except they are based on correlations between scores for each task and Rasch measures of examinee ability (instead of total raw scores). Greater point-measure correlations suggest greater sensitivity to the construct being targeted by the task.

Stage 4: Investigating the nature of engagement. While the results from Stage 3 explain the overall effects of certain task characteristics on performance, Stage 4 involved disentangling the direct effects of task characteristics on performance from their effects on engagement. To accomplish this, structural models of task engagement and performance were specified. Structural equation modeling involves performing a series of regression equations simultaneously to test hypothesized relationships among various observed and latent variables (Kline, 2016). More specifically, multiple-indicators multiple-causes (MIMIC) models (Joreskog & Goldberger, 1975) were specified and compared in order to investigate the relationships between task characteristics, engagement, and performance as well as the stability of those relationships across examinee age groups. Stage 4 resulted in additional evidence about the meaning of scores in the context of an explanation inference in a validity argument.

To begin, the results of Stage 3 were reframed as a factor model at the task level. First, a model of how each performance was indicated by two rater scores was specified (see Figure 3.1). Each rater score was also influenced by a measure of rater severity and by the rubric type (0-to-3 points or 0-to-5 points) that was employed for the task. The effects of rater severity and the rubric type on each of the scores assigned were constrained to be equal.

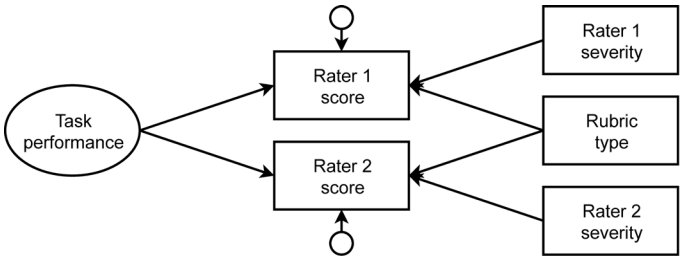


Figure 3.1. Model of rater scores.

The model was then expanded into a MIMIC model to include "causes" or contributors to task performance, such as personal attributes and task characteristics (see Figure 3.2). In this model, task performance was treated as the product of the interaction between personal attributes (represented in the top left) and the task characteristics (represented in the lower left). The personal attribute that was most expected to affect performance was *examinee ability*. (Examinee age, another personal attribute, would be addressed separately later.) *Task type difficulty* (calculated in Stage 3) and dummy variables (i.e., 1s and 0s) for the presence or absence of *topical choice*, *vocabulary support*, *novelty*, and *video animation* were treated as task characteristics that also contributed to performance.

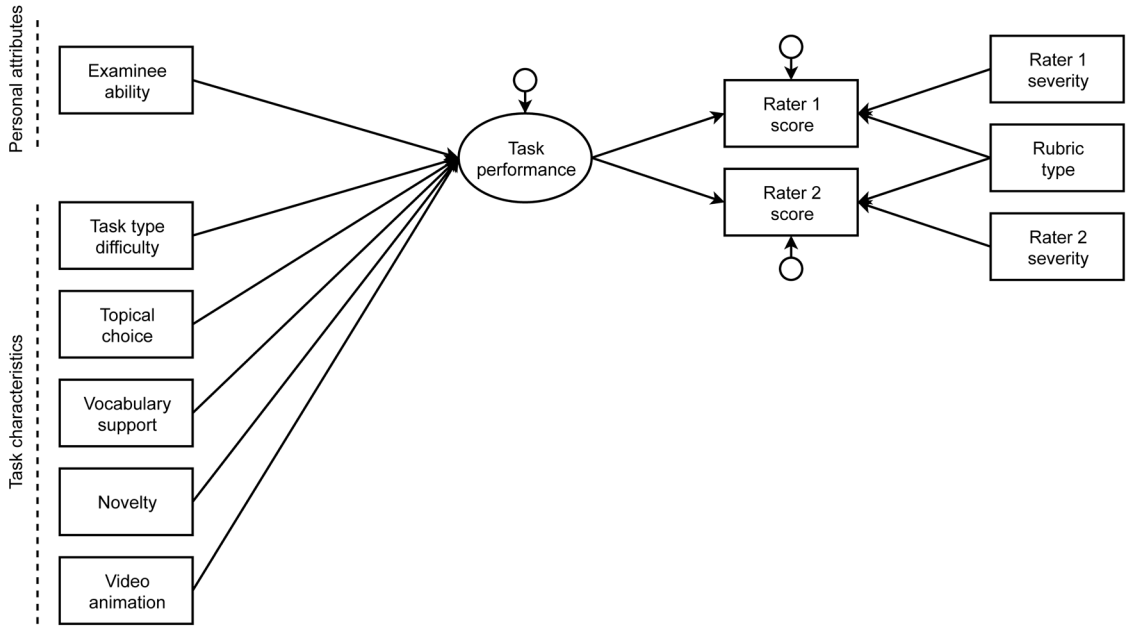


Figure 3.2. MIMIC model of task performance.

Since the presence of topical choice, vocabulary support, novelty, and video animation were also hypothesized to support task engagement, the model of task performance (from Figure 3.2) was expanded to include task engagement, as seen in Figure 3.3. Task engagement was thus indicated by the three acoustic features identified as likely indicators of engagement. These included the mean harmonicity, the mean shimmer, and the standard deviation of shimmer. Because the model in Figure 3.3 was considered the baseline model, no path was drawn between task performance and task engagement, suggesting that no structural relationship exists between the two.

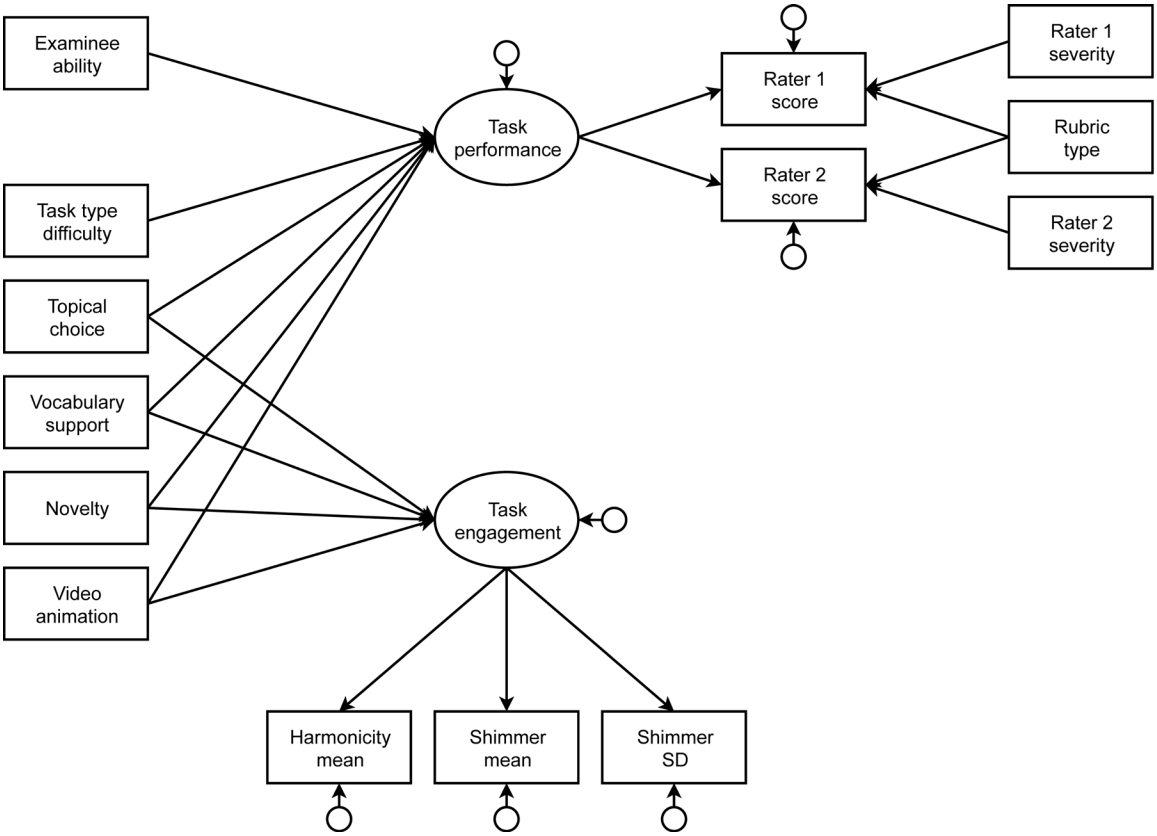


Figure 3.3. Baseline MIMIC model of task performance and engagement.

A correlational relationship between task performance and task engagement was anticipated, so three additional models were specified for comparison to the baseline model and with each other. Figure 3.4 shows the hypothesized model. Task engagement was predicted to

have an effect on task performance, so a path from *task engagement* to *task performance* was included in the hypothesized model. However, two other possibilities—that task performance affects task engagement and that task performance and task engagement affect each other—were also considered. As a result, an alternate model with a path from *task performance* to *task engagement* (see Figure 3.5) and a second one with paths in both directions (see Figure 3.6) were specified.

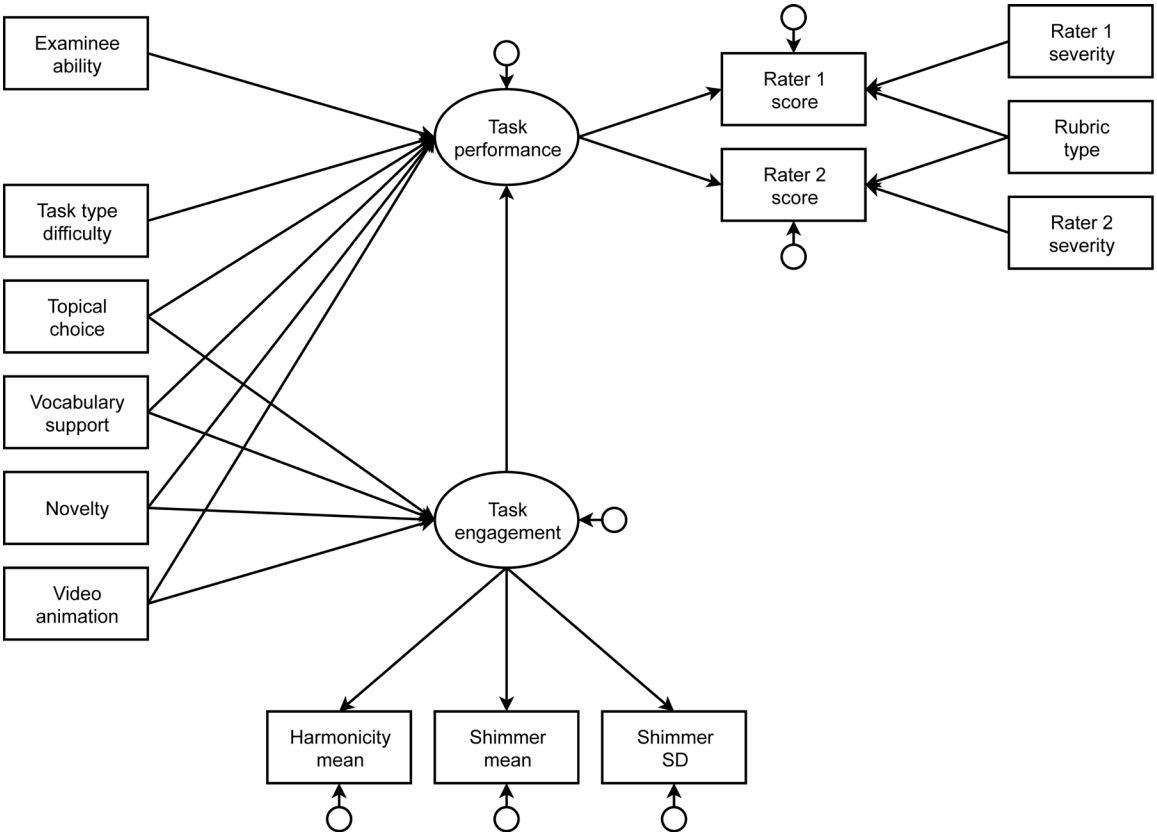


Figure 3.4. Hypothesized MIMIC model of task performance and engagement.

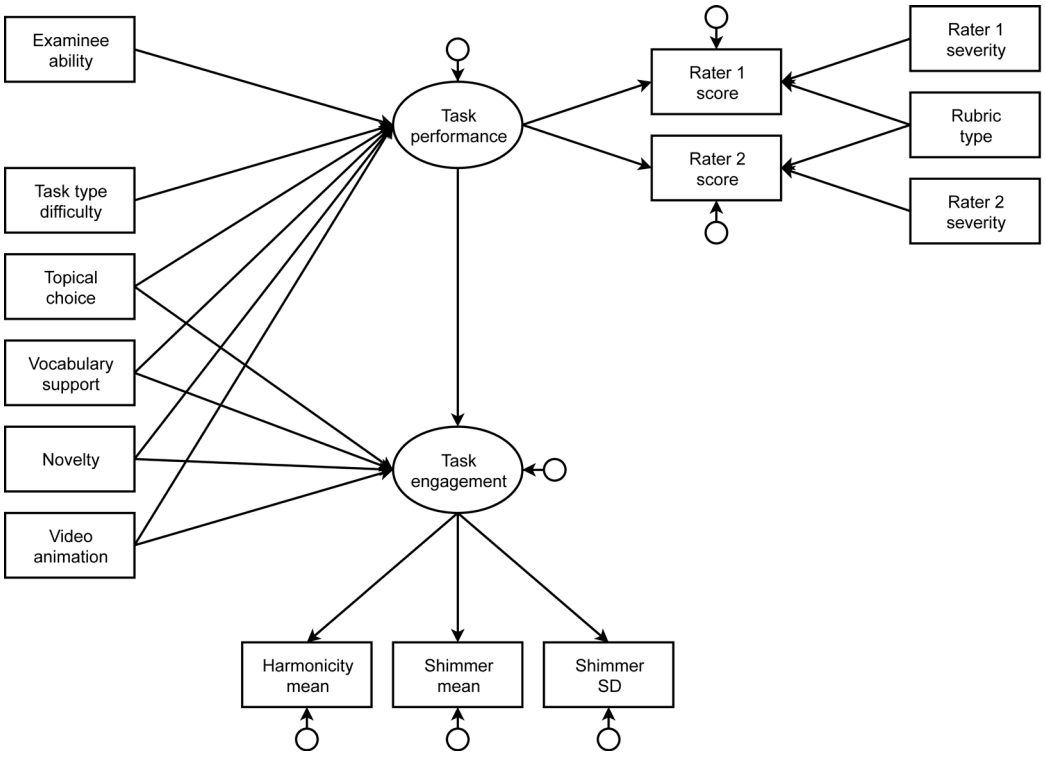


Figure 3.5. First alternate MIMIC model of task performance and engagement.

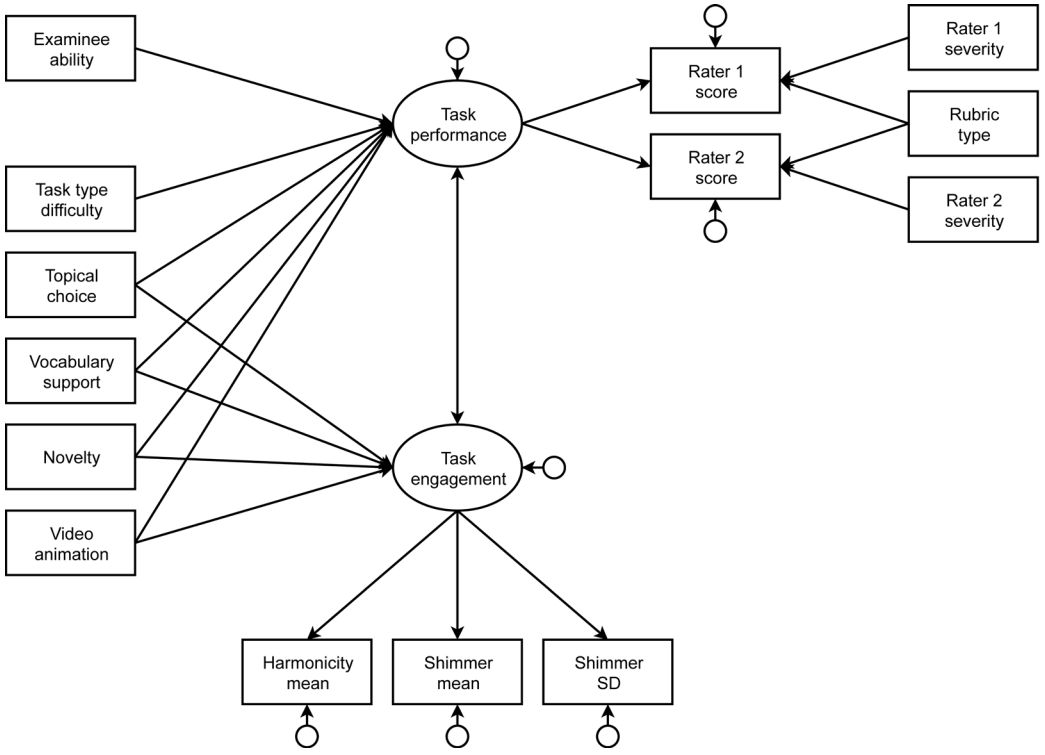


Figure 3.6. Second alternate MIMIC model of task performance and engagement.

All four models were specified in Mplus (Version 8; Muthén & Muthén, 2017) with cases clustered by examinee. Model fit was evaluated based on chi-square statistics and common indices listed in Table 3.3. (For an explanation and discussion of the various indices, consult Hu and Bentler, 1999.) The best fitting model was retained, and then nonsignificant paths were tested to see whether they should be part of the final model. This involved re-specifying the model with the paths removed one at a time and then applying chi-square difference tests to clarify whether the presence of each path resulted in a statistically significant improvement in model fit. From the final model, direct and indirect effects of each task characteristic on performance could be calculated.

Table 3.3

Indicators of Good Model Fit

Index	Shorthand	Expected Value
Root mean square error of approximation	RMSEA	< 0.06
Comparative fit index	CFI	> 0.95
Tucker-Lewis fit index	TLI	> 0.95
Standardized root mean square residual	SRMR	< 0.08

Note. These values are adopted from Hu and Bentler (1999)

Lastly, the structural invariance of the final model across age groups was evaluated in a series of steps (see Bollen, 1989). Each step involved testing a hypothesis that is nested in the one preceding it. The hypothesis that the form of the model fits both the oldest and youngest third of participants well was tested in the first step. To accomplish this, a configural model was specified in which latent factor means were constrained at zero but factor loadings and intercepts were freely estimated for each group. The second step involved testing the hypothesis that the slopes relating the indicators to the latent variables were consistent between groups. Therefore, a metric model was specified in which factor loadings were constrained to be equal between groups. The metric model was then compared to the configural model using a chi-square difference test. In the third step, the hypothesis that intercepts were consistent between groups was tested by specifying

a scalar model, with intercepts constrained to be equal between groups, and comparing it to the metric model using a chi-square difference test. After that, the hypothesis that factor means were invariant across groups was tested. The factor invariance model was again compared to the preceding model using a chi-square difference test. Lastly, the hypothesis that residual error variances were invariant between groups was tested. As before, a chi-square difference test was used to confirm the hypothesis.

The Qualitative Phase

The purpose of the second phase of the study was to investigate task engagement among YLLs using a qualitative approach. After a subset of *TOEFL Primary* Speaking test tasks were administered to YLLs, retrospective structured interviews were conducted to inquire about their experiences with the tasks and the task characteristics of topical choice, vocabulary support, novelty, and video animation. Information gleaned from verbal reports confirmed and expanded upon findings from Stage 4 about task characteristics that support engagement.

Participants

Eight young English-as-a-foreign-language learners living in Daejeon, South Korea, were recruited to participate in this phase of the study. The participants were members of the population targeted by the *TOEFL Primary* Speaking test, and they were selected based on a convenience sampling plan. The five boys and three girls were all eight or nine years old and spoke Korean as their first language.

Instruments

A subset of four *TOEFL Primary* Speaking test tasks from Stage 1 was administered (see Table 3.4). These tasks were selected because collectively they contained all the task characteristics being investigated in the current study.

Table 3.4

Tasks Administered in the Qualitative Phase

Task	Task Name	Task Type	Task Characteristics
1	What's your favorite animal?	Express an Opinion	Topical choice, vocabulary support
3	What's strange on the bus?	Describe a Picture	Vocabulary support, novelty
5	What happened to the key?	Retell a Story	Vocabulary support, novelty, video animation
7	Ask three questions about the tiger.	Ask Questions	Topical choice

Procedures

Each participant individually completed the test tasks immediately before being interviewed about his or her experience with the tasks and the task characteristics. An introduction to the activity, which was adapted from Johnstone, Bottsford-Miller, and Thompson (2006), was first read to each participant. The participants were then asked if they would be willing to take a test and then answer questions about it while being videotaped. After the test concluded, a conversation ensued to elicit information about what task the participant found to be most engaging and how different task characteristics made the tasks more (or less) interesting and enjoyable. A script containing the introductory explanation and the follow-up questions can be found in Appendix E. A translator was present throughout the interviews, so interview questions were asked and answered in English or Korean or a mix of the two languages according to each participant's preference.

Data Analysis

Responses to the interview questions were transcribed (and translated as needed), and the transcriptions were entered into NVivo (Versions 12; QSR International, 2018) for analysis. The transcriptions were then tagged by topic so that patterns in the responses could be examined. The response patterns helped shed light on the relationships between task characteristics and task engagement identified in Stage 4. In addition, the response patterns suggested some directions for future research.

Summary

In Chapter 3, a description of the methodology for the current study was described. A variety of quantitative and qualitative approaches were employed to address the research questions. In addition to providing evidence for the validity of a speaking assessment for YLLs, the current study explored potential relationships between examinee age, task characteristics, task engagement, and task performance. The findings are presented in the next chapter.

Chapter IV

RESULTS

Chapter 4 details the results of the quantitative (i.e., Stages 1 through 4) and qualitative analyses outlined in Chapter 3. The results of Stage 1 describe how effective *TOEFL Primary* Speaking tasks, raters, and rating scales were for evaluating young learner speaking abilities. The generalizability of scores over parallel versions of tasks and test forms and across raters and examinee age groups was revealed in Stage 2. Stage 3 demonstrated how the task characteristics of topical choice, vocabulary support, novelty, and video animation systematically influenced task difficulty. The structural relationships among these task characteristics, task performance, and task engagement were explored in Stage 4. The qualitative phase expanded upon the quantitative findings with YLL self-reports about their experiences with the test tasks and their characteristics.

Evidence That Scores Reflect Underlying Speaking Abilities

Before evaluating the scores assigned to the responses, rater consistency was examined. The exact agreement rate between raters was 74.0% for tasks scored from 0 to 3 and 59.4% for tasks scored from 0 to 5. With variations in the final composition of raters for each response, Krippendorff's Alpha (Hayes & Krippendorff, 2007) was calculated as a measure of interrater reliability. The resulting statistic was 0.87. With the subsequent development of training and calibration materials, accuracy rates for scoring the *TOEFL Primary* Speaking test have likely improved since this first attempt to score the then-new test.

The following descriptive statistics are based on the average (mean) of the two ratings assigned to each response (k=11). Total scores ranged from 1.5 to 40.5 points, with a mean of

23.41 points out of a possible 41 points. The standard deviation was 10.33 points. The skewness and kurtosis values for total scores were –0.31 and –1.11, respectively—within acceptable limits for the data to be considered normally distributed (Kline, 2016). Since tasks were the object of interest in this study, descriptive statistics for each of the 11 scored tasks are presented in Table 4.1. In addition, an internal consistency reliability coefficient in the form of Cronbach's alpha was calculated to be 0.97.

Table 4.1

Descriptive Statistics for Individual Tasks (*n*=401)

Task	Task Type	Point Range	Point Mean	Percent Mean	Point S.D.	Skewness	Kurtosis
1	Express an Opinion	0 to 3	2.22	74.0	0.80	–0.78	–0.37
2	Give Directions	0 to 5	2.26	45.2	1.57	–0.15	–1.30
3	Describe a Picture	0 to 3	1.87	62.3	0.81	–0.16	–0.91
4	Describe a Picture	0 to 3	2.06	68.7	0.75	–0.89	0.40
5	Retell a Story	0 to 5	2.35	47.0	1.24	0.30	–0.68
6	Ask Questions	0 to 3	1.81	60.3	1.11	–0.32	–1.34
7	Ask Questions	0 to 3	1.88	62.7	0.89	–0.52	–0.61
8	Retell a Story	0 to 5	2.49	49.8	1.26	–0.40	–0.66
9	Ask Questions	0 to 3	1.76	58.7	1.04	–0.11	–1.34
10	Give Directions	0 to 5	2.63	52.6	1.47	–0.38	–0.95
11	Express an Opinion	0 to 3	2.07	69.0	0.88	–0.79	–0.20

In order to gather further evidence to support assumptions underlying the evaluation inference in a validity argument, Rasch analyses were performed to examine how well the scores reflected the underlying abilities of YLLs, how effectively rating scales were applied by raters, how consistently tasks functioned across age groups, and how consistently rating scales were applied by raters across age groups. About 5% (4.4%) of observations had standardized residuals outside ±2 and about 1% (0.7%) outside ±3, indicating that the data fit the Rasch model well (Linacre, 2017).

Figure 4.1 visually depicts the Rasch measures for examinee ability, task difficulty, and rater severity. The first column (Measure) represents a logit scale, which is an equal-interval scale

upon which examinee ability, task difficulty, and rater severity are all mapped. The second column shows the ability estimates of the examinees. Higher-ability examinees are represented at the top of the column, and lower-ability ones at the bottom. Each asterisk represents four examinees. The examinee ability mean was 0.84 logits with a standard deviation of 2.49, and ability estimates ranged 12.36 logits, spanning from –4.98 logits to 7.38 logits. A chi-square test indicated that examinees significantly differed in ability, X^2 (400) = 14366.9, $p < 0.01$. Furthermore, a separation index of 6.42 reflected over eight statistically distinct strata of examinees (cf. Wright & Masters, 1982), and the reliability of the separation of examinees into these strata was high at 0.98. These indicators of the magnitude of differences in ability estimates among examinees suggest that the *TOEFL Primary* Speaking tasks were sufficiently able to distinguish examinees across the spectrum of ability levels observed in the study.

Measure	Examinees	Tasks		Raters	0-to-3 R.S.	0-to-5 R.S.
8	More able	More difficult		More severe	3	5
	.					
7						
	.					
6	.					
	*.					
5	*.					
	****.					———
4	****					
	*****.					
3	******				———	4
	*******.					
2	******.					
	*********	2_Give_Directions	5_Retell_a_Story		2	———
1	*******.	8_Retell_a_Story				
	****.	10_Give_Directions		C		3
0	******.	6_Ask_Questions	9_Ask_Questions	B D E F H I	———	
	*****.	3_Describe_a_Picture	7_Ask_Questions	A G		———
–1	****.	4_Describe_a_Picture	11_Express_an_Opinion			2
	*****.	1_Express_an_Opinion			1	
–2	******.					———
	*****.					
–3	***.				———	1
	.					
–4	*.				0	———
	*					0
–5	.					
	Less able	Less difficult		Less severe		

Figure 4.1. Variable map.

The third column depicts the difficulty estimates for each task. Task fit statistics—both the weighted infit mean square and the unweighted outfit mean square—fell within two standard

deviations of the mean, providing further evidence of model fit (cf. Pollitt & Hutchinson, 1987). The task difficulty mean was centered at 0 logits, and the standard deviation was 0.96 logits. The tasks spanned a range of difficulty, from a relatively easy –1.47 logits for the first *Express an Opinion* task to a more difficult 1.45 logits for a *Give Directions* task. The *Give Directions* and *Retell a Story* task types were the most difficult, and the *Express an Opinion* and *Describe a Picture* task types were the easiest. The relationship between the distributions of examinee ability and task difficulty estimates suggests that the tasks could be performed with some level of success by most test takers. The opportunity to experience success is especially desirable in a test for young learners. However, the range of task difficulty was much smaller than the range of examinee abilities, which indicates redundancy in the difficulty levels of the tasks. The cost of this redundancy was a limitation in the sensitivity of the instrument to the lowest and highest ability levels. The inclusion of a few easier and more difficult tasks to better cover the full spectrum of abilities could easily remedy this.

The fourth column in Figure 4.1 shows the severity of the raters. More severe raters are represented toward the top of the column, and more lenient ones toward the bottom. The rater severity mean was centered at 0 logits with a standard deviation of 0.22, and the range of severity was less than one logit, ranging from –0.32 to 0.30 logits. According to Myford and Wolfe (2000), the range of rater severity (0.62 logits) should be less than half the range of examinee ability estimates (12.36 / 2 = 6.18 logits), and here, it was substantially less. Such little variation in rater severity along with no misfitting rater statistics suggests that the rating scales were consistently applied by raters.

More evidence about the functioning of the rating scales described by the rubrics can be seen in the variable map's last two columns, which show the thresholds for the score categories for each rating scale. A horizontal line between score categories represents the threshold at which test takers have an equal chance of getting either score, and each category shows the score

examinees at a certain ability level would most likely receive. For example, a horizontal line drawn from an examinee whose ability estimate was –1 logit would fall within the score categories of 1 on the 0-to-3-point rating scale and 2 on the 0-to-5-point rating scale, indicating the most probable scores received for responses by that examinee. The probability curves for the rubrics' rating scales (see Figures 4.2 and 4.3) make it clear that the score categories are distinct and well distributed across the logit range.

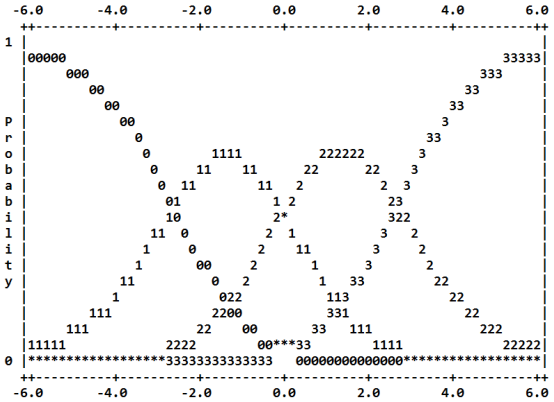


Figure 4.2. Probability curves for 0-to-3-point rating scale.

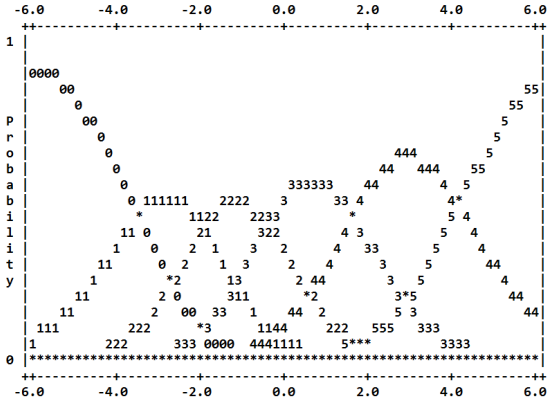


Figure 4.3. Probability curves for 0-to-5-point rating scale.

More detailed information about rating scale functioning can be gleaned from the category statistics (see Tables 4.2 and 4.3). The first three columns of each table show the score category, the number of observations in each category, and the percentage of observations in each category. The scores appear reasonably distributed across the categories. The fourth column

shows fit statistics. Outfit mean squares less than 2 indicate relatively low levels of noise in the data. The next column describes the step calibrations, which are the examinee ability measures (along with their standard errors) at which the probability of responses falling into each score category begins to exceed the probability of responses falling into the previous score category. These step calibrations correspond to the thresholds drawn between score categories in the last two columns of Figure 4.1 and to the points at which the probability curves intersect in Figures 4.2 and 4.3. The distance between advancing step calibrations should optimally be between 1.4 and 5 logits (Linacre, 2002), and that is the case with both rating scales, suggesting that the numbers of score categories used were appropriate. The last column shows the average ability measures expected for each score category. Monotonocity is clearly demonstrated by the ordered progression in ability estimates from one score category to the next. Both rubrics and their associated rating scales appear to have functioned effectively.

Table 4.2

Category Statistics for 0-to-3-Point Rating Scale

Score Category	Count	%	Outfit Mean Square	Step Calibrations Measure	S.E.	Expected Measure
0	455	8%	1.0			–3.75
1	1306	23%	1.0	–2.63	0.06	–1.33
2	1896	34%	1.2	–0.03	0.04	1.33
3	1957	35%	1.1	2.66	0.04	3.78

Table 4.3

Category Statistics for 0-to-5-Point Rating Scale

Score Category	Count	%	Outfit Mean Square	Step Calibrations Measure	S.E.	Expected Measure
0	389	12%	1.0			–4.68
1	528	16%	0.7	–3.64	0.07	–2.78
2	626	20%	0.8	–1.93	0.06	–1.13
3	824	26%	0.8	–0.34	0.06	0.66
4	655	20%	0.9	1.66	0.06	2.93
5	186	6%	1.1	4.24	0.09	5.31

Bias analyses were then specified to investigate possible task–age group and rater–age group interactions. By dividing the resulting bias measures by the standard errors, *z* scores were calculated. Any *z* scores outside ±2 would indicate the presence of significant bias or interaction (McNamara, 1996). For most tasks, no evidence of significant bias related to age was observed. This indicates that tasks tended to function similarly for all examinees regardless of age. This finding supports the assumption that tasks function consistently across age groups. The results of the bias analyses can be seen in Table 4.4. The observed – expected average shows the difference between observed measures and the measures predicted by the Rasch model for each interaction. The bias measure, in logits, shows how much more difficult (negative values) or easy (positive values) the task was for each group.

Table 4.4

Task–Age Group Bias Analyses

Task	Task Type	Age Group	Observed – Expected Average	Bias Measure	Model S.E.	*Z* Score
1	Express an Opinion	Youngest	–0.09	0.31	0.11	2.74
2	Give Directions	Youngest	0.03	–0.06	0.09	–0.70
3	Describe a Picture	Youngest	–0.01	0.04	0.11	0.36
4	Describe a Picture	Youngest	0.05	–0.15	0.11	–1.34
5	Retell a Story	Youngest	0.03	–0.05	0.09	–0.60
6	Ask Questions	Youngest	0.00	–0.02	0.11	–0.14
7	Ask Questions	Youngest	–0.04	0.11	0.11	1.02
8	Retell a Story	Youngest	0.05	–0.09	0.09	–1.06
9	Ask Questions	Youngest	0.06	–0.19	0.11	–1.74
10	Give Directions	Youngest	–0.01	0.02	0.09	0.25
11	Express an Opinion	Youngest	–0.06	0.21	0.11	1.85
1	Express an Opinion	Oldest	0.06	–0.24	0.13	–1.90
2	Give Directions	Oldest	–0.06	0.12	0.09	1.31
3	Describe a Picture	Oldest	0.10	–0.34	0.12	–2.89
4	Describe a Picture	Oldest	0.00	0.00	0.12	0.03
5	Retell a Story	Oldest	–0.02	0.05	0.09	0.51
6	Ask Questions	Oldest	0.02	–0.06	0.12	–0.51
7	Ask Questions	Oldest	–0.02	0.05	0.12	0.46
8	Retell a Story	Oldest	–0.02	0.04	0.09	0.43
9	Ask Questions	Oldest	–0.06	0.21	0.11	1.84
10	Give Directions	Oldest	–0.02	0.03	0.09	0.35
11	Express an Opinion	Oldest	0.02	–0.07	0.12	–0.57

The two exceptions were small but statistically significant interactions between the youngest age group and Task 1 and the oldest age group and Task 3. Task 1 was an *Express an Opinion* task that featured topical choice and vocabulary support, and the youngest students did slightly more poorly than the Rasch model predicted. Task 3 was a *Describe a Picture* task that featured vocabulary support and novelty, and the oldest examinees did slightly better than predicted by the Rasch model. These were isolated findings in that other tasks of the same types or with the same task characteristics did not exhibit similar interactions. One possibility that might explain both situations is that vocabulary support subtly adds to the cognitive processing required by the tasks, disadvantaging the youngest students. Another possibility that might explain the interaction with Task 1 is that, developmentally, facility with expressing and explaining opinions on certain topics may develop later than the other skills tested. For the interaction with Task 3, it may be the case that older students have more experience describing pictures or at least the contents of certain pictures. However, the differences between the observed and expected averages were ±0.10 logits or less, which "are too small to affect individual student measures. The results may be more usefully explained as…differences than…bias" (Du & Wright, 1997, p. 18).

For all raters, no evidence of significant bias related to age was observed. This means that rater behavior was generally not influenced by examinee age. This finding supports the assumption that raters can apply the rubrics consistently to responses from examinees of varying ages. The results of the bias analyses can be seen in Table 4.5. The observed – expected average shows the difference between observed measures and the measures predicted by the Rasch model within each age group. The bias measure, in logits, shows how much more lenient (negative values) or severe (positive values) the rater was for each group.

Table 4.5

Rater–Age Group Bias Analyses

Rater	Age Group	Observed – Expected Average	Bias Measure	Model S.E.	*Z* Score
A	Youngest	0.01	–0.02	0.06	–0.32
B	Youngest	–0.01	0.03	0.12	–0.25
C	Youngest	0.06	–0.21	0.15	–1.42
D	Youngest	0.03	–0.10	0.08	–1.27
E	Youngest	0.00	0.00	0.09	0.00
F	Youngest	–0.03	0.07	0.15	0.48
G	Youngest	–0.05	0.15	0.11	1.35
H	Youngest	0.02	–0.06	0.11	–0.56
I	Youngest	–0.06	0.19	0.10	1.95
A	Oldest	0.00	–0.01	0.07	–0.16
B	Oldest	–0.02	0.04	0.13	0.27
C	Oldest	–0.01	0.05	0.16	0.28
D	Oldest	0.01	–0.04	0.07	–0.57
E	Oldest	0.00	0.00	0.10	–0.02
F	Oldest	–0.10	0.28	0.15	1.85
G	Oldest	0.01	–0.05	0.11	–0.42
H	Oldest	–0.04	0.10	0.11	0.94
I	Oldest	0.03	–0.10	0.10	–1.04

With over 400 examinees, a great deal of variation in terms of examinee ability was measured by the test. Test tasks also demonstrated reasonable and appropriate variations in difficulty. Raters showed little variability during constructed-response scoring, and the rating scales separated responses into distinct, monotonic categories. In terms of rater bias, no systematic interactions between raters and examinees grouped by age were uncovered. These findings support assumptions that underlie the claim that observations of performance (i.e., the scores) have the intended characteristics.

The Consistency of Scores Across Tasks, Ratings, and Forms

In order to gather evidence to support assumptions underlying the generalization claim of a validity argument, a series of generalizability and decision studies (G- and D-studies) were conducted. A G-study was employed to estimate the percentages of score variance that were

attributable to examinees, tasks, and ratings. D-studies revealed the effects of varying the number of ratings and the number of tasks of different types on overall score dependability. Lastly, the relative contributions of different task types to composite score variance across age groups were evaluated.

The results of the G-study are presented in Table 4.6. The percentage of score variance accounted for by examinees (persons) on individual tasks ranged from 58% for the *Describe a Picture* tasks to 79% for the *Give Directions* tasks. Furthermore, the percentage of score variance attributable to tasks or ratings was 3% or less across all tasks. A high percentage of variance attributable to examinees and not tasks or ratings is desirable, suggesting that scores primarily reflect variations among examinees and not tasks or ratings. Nonetheless, there is also some evidence of interaction, especially between examinees and tasks. Of course, some variability here would be expected as tasks were designed to have some construct-relevant variation, and in no case did the percentage of score variance exceed 25%. The percentages of score variance accounted for by interactions between tasks and ratings and between examinees and ratings were very low across all task types, being at or near 0%. There was also some score variance attributable to the three-way interactions between examinees, tasks, and ratings coupled with error. For example, these interactions and error accounted for 23% of the score variance with the *Describe a Picture* tasks. This is likely due to the fact that vocabulary support was only available for one of the two tasks. While not desirable, the error variance is understandable given that it was such a new test. As task specifications become more refined, these percentages can be expected to drop. Overall, however, the greatest proportions of variance were attributable to examinees and, to a lesser extent, their interactions with tasks. This finding further supports the assumption that scores are consistent between parallel tasks and ratings.

Table 4.6

Estimated Variance Components for Individual Tasks of Different Types

Source of Variation	Give Directions		Retell a Story		Ask Questions		Describe a Picture		Express an Opinion	
Persons (*p*)	1.96	79%	1.32	76%	0.76	70%	0.42	58%	0.47	61%
Tasks (*t*)	0.06	3%	0.00	0%	0.00	0%	0.02	2%	0.01	1%
Ratings (*r′*)	0.00	0%	0.02	1%	0.00	0%	0.00	0%	0.00	0%
pt	0.27	11%	0.15	8%	0.20	19%	0.11	16%	0.18	24%
pr′	0.01	0%	0.00	0%	0.01	0%	0.00	0%	0.01	1%
tr′	0.00	0%	0.01	1%	0.00	0%	0.00	0%	0.00	0%
ptr′,*e*	0.16	7%	0.24	14%	0.12	11%	0.17	23%	0.10	13%

To examine the effects of varying the configuration of the test in terms of the numbers of ratings and tasks of different types on score dependability, a series of decision studies was conducted. First, the effects of keeping the number of tasks (i.e., one task from each task type) constant while increasing the numbers of ratings per task were compared to the effects of keeping the number of ratings per task (i.e., one rating) constant while increasing the number of tasks. Figure 4.4 shows how these changes to the test configuration, while maintaining the same overall number of ratings between the two conditions, affect the ϕ dependability coefficient. Apparently, increasing the number of tasks with a single rating has a greater effect on ϕ than increasing the number of ratings per task. Considering that task development is a one-time expense while scoring becomes a recurring one, increasing the number of tasks can also be viewed as more practical, especially when considering how more tasks can result in more opportunities to measure the full range of examinee ability levels. These findings about the benefits of adjusting the number of tasks instead of ratings are consistent with the results from previous research on the *TOEFL Primary* Speaking test's big brother, the *TOEFL* iBT (Lee, 2005, 2006).

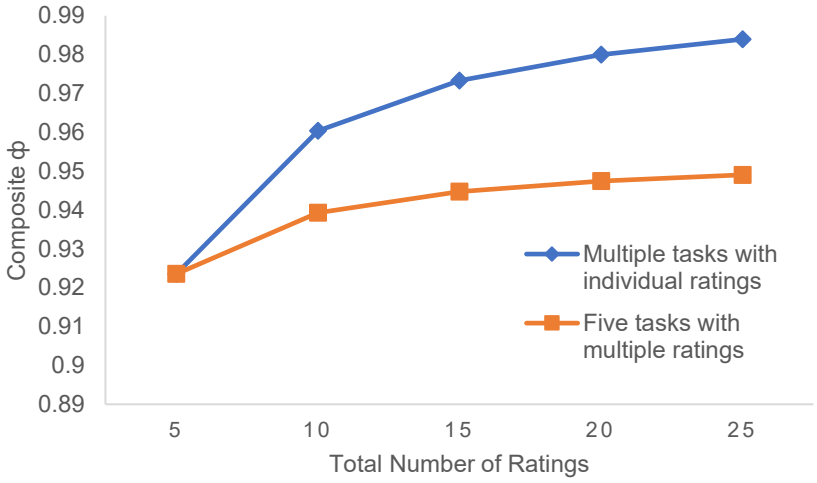


Figure 4.4. Effects of varying the numbers of tasks and ratings.

A second series of decision studies investigated the effect of increasing the number of tasks on a test form on overall score dependability. The hypothetical baseline was one task for each task type (i.e., five tasks in all). Table 4.7 shows all of the possible configurations, from five tasks to ten tasks, with one or two tasks of each task type to ensure broad construct representation. The rows are sorted first by the number of tasks and then by the composite ϕ, an indicator of score dependability. As the number of tasks increases, the improvements in ϕ diminish. Considering how test length may be a concern for YLLs with more limited attentional resources (not to mention costs related to development, administration, and scoring), seven appears to be the optimal number of tasks for maximizing score dependability. The most dependable configurations with seven tasks all contain an additional task scored 0 to 5 (i.e., a *Give Directions* or *Retell a Story* task). Although configuration 7A has the highest composite ϕ, with two additional tasks scored 0 to 5, which tend to be longer than those scored 0 to 3, it is most likely not ideal after considering concerns about test length against the limited gains in score dependability. Therefore, configuration 7B is the recommended configuration for maximizing score dependability. This configuration, along with configuration 7D, is how operational forms of the *TOEFL Primary* Speaking test are actually configured (Cho et al., 2016, 2017).

Table 4.7

Dependability of Hypothetical Test Form Configurations

Configuration	Number of Tasks of Each Task Type					Total Tasks		ϕ
	Give Directions	Retell a Story	Ask Questions	Describe a Picture	Express an Opinion	Scored 0 to 5	Scored 0 to 3	
5	1	1	1	1	1	2	3	0.92
6A	2	1	1	1	1	3	3	0.94
6B	1	2	1	1	1	3	3	0.94
6C	1	1	2	1	1	2	4	0.94
6D	1	1	1	2	1	2	4	0.93
6E	1	1	1	1	2	2	4	0.93
7A	2	2	1	1	1	4	3	0.95
7B	2	1	2	1	1	3	4	0.95
7C	2	1	1	2	1	3	4	0.95
7D	1	2	2	1	1	3	4	0.95
7E	2	1	1	1	2	3	4	0.95
7F	1	2	1	2	1	3	4	0.94
7G	1	2	1	1	2	3	4	0.94
7H	1	1	2	2	1	2	5	0.94
7I	1	1	2	1	2	2	5	0.94
7J	1	1	1	2	2	2	5	0.94
8A	2	2	2	1	1	4	4	0.95
8B	2	2	1	2	1	4	4	0.95
8C	2	2	1	1	2	4	4	0.95
8D	2	1	2	2	1	3	5	0.95
8E	2	1	2	1	2	3	5	0.95
8F	1	2	2	2	1	3	5	0.95
8G	2	1	1	2	2	3	5	0.95
8H	1	2	2	1	2	3	5	0.95
8I	1	2	1	2	2	3	5	0.95
8J	1	1	2	2	2	2	6	0.95
9A	2	2	2	2	1	4	5	0.96
9B	2	2	2	1	2	4	5	0.96
9C	2	1	2	2	2	3	6	0.96
9D	2	2	1	2	2	4	5	0.96
9E	1	2	2	2	2	3	6	0.95
10	2	2	2	2	2	4	6	0.96

Based on the optimal test configuration identified above (i.e., configuration 7B), the effective weights of each task type to the composite universe score variance according to examinee age groups are presented in Table 4.8. The pattern of contributions from each task type was remarkably consistent across age groups, further indicating that the roles of the different task types in the overall generalizability of scores from one form to the next were similar regardless of examinee age.

Table 4.8

Contributions to Composite Universe Score Variance Across Age Groups

Group	Give Directions	Retell a Story	Ask Questions	Describe a Picture	Express an Opinion
Youngest	39.45%	15.97%	25.47%	9.11%	10.01%
Oldest	39.60%	16.25%	24.72%	9.72%	9.71%

In summary, the bulk of score variance on *TOEFL Primary* Speaking test tasks across all five task types was attributable to examinees. Decision studies revealed that the reliability of scores from one form to another was maximized by increasing the number of tasks scored on a scale of 0 to 5 (i.e., the *Give Directions* or *Retell a Story* tasks) and not by increasing the number of ratings. Finally, the relative contributions of different task types to the dependability of scores were consistent across age groups. These findings support assumptions that underlie the claim that scores are estimates of expected scores over relevant parallel versions of tasks and test forms and across ratings and examinee age groups.

The Influence of Speaking Task Characteristics on Performance

In Stage 3, the Rasch analysis conducted in Stage 1 was expanded to distinguish difficulty estimates attributable to individual task characteristics through the application of Fischer's (1973, 1995) linear logistic test model (LLTM). Doing so provided support for the assumption that underlies the explanation inference in a validity argument. More specifically, the

results of the analyses function as evidence to show how task characteristics systematically influence task difficulty as predicted by the simplified model of speaking performance (see Figure 2.6).

Data that fit the Rasch model in Stage 1 also fit LLTM quite well, with the distributions of score residuals and infit mean square statistics for examinees, the decomposed tasks (i.e., the task characteristics that make up the tasks), and raters all falling within two standard deviations of the mean (cf. Pollitt & Hutchinson, 1987). In addition, the reliability of separation indices for difficulty estimates attributable to the task type, topical choice, vocabulary support, novelty, and video animation were 1.00, 0.97, 0.99, 1.00, and 0.88, respectively. These high values indicate that the influences of these task characteristics on overall task difficulty and, by extension, task performance were distinct and quantifiable.

The difficulty estimates for each task type and the presence or absence of topical choice, vocabulary support, novelty, and video animation are depicted in Figure 4.5. Measures for examinee ability and rater severity were similar to those found in Stage 1, so they are not shown here. The first column (Measure) is the logit scale used to measure the difficulty levels of the task characteristics, which are centered around 0. Greater positive values indicate increasing difficulty, and negative values indicate increasing facility. Difficulty estimates for the five task types range from –1.37 for the easy *Describe a Picture* tasks to 1.77 for the difficult *Give Directions* tasks. Tasks that contained the element of topical choice were approximately 0.37 logits easier than tasks that did not contain it. One could easily see how being able to choose a familiar topic to talk about could improve performance. However, tasks with vocabulary support, novelty, and video animation were about 0.65, 0.99, and 0.22 logits more difficult than tasks that did not contain them, respectively. Vocabulary support was hypothesized to support performance because it would help examinees with language that could be useful to completing a task instead of them getting stuck by not knowing a few key words. The evidence suggests, however, that vocabulary

support did not really help test takers. Instead, it may have served as additional stimulus material, thus increasing the cognitive demands of a task. Likewise, novelty was expected to support performance, but it actually had the strongest negative influence on performance of the task characteristics examined. Perhaps novelty served to enhance the linguistic demands of tasks because it obligated novel language production—meaning that it was likely construct relevant. And lastly, video animation appears to have had a negative influence on performance, albeit small. However, with only one task that featured this task characteristic, it is difficult to draw any firm conclusions from this finding.

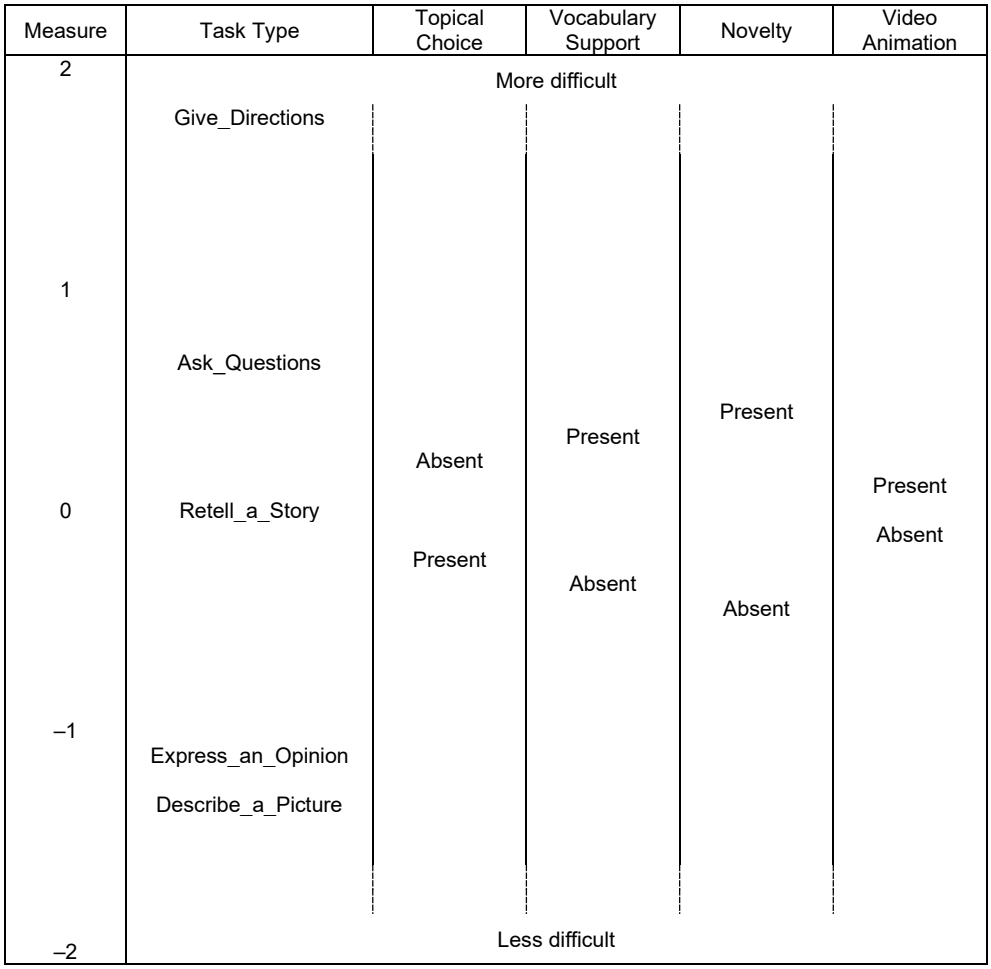


Figure 4.5. Map of difficulty estimates for component task characteristics.

These findings paint an incomplete picture about the effects of these task characteristics on performance. For example, it is not clear how much of the positive influence of topical choice on performance is because topical choice supports examinee engagement, which in turn supports performance, and how much is because topical choice directly supports performance. Also, the fact that vocabulary support, novelty, and video animation all had negative effects on performance poses a challenge to assumptions that these task characteristics would support both engagement and performance. These issues were more closely examined in Stage 4.

Table 4.9 summarizes the effects of topical choice, vocabulary support, novelty, and video animation on qualities of measurement. In addition to task difficulty determined by LLTM, measurement qualities as reflected by discrimination indices and point–measure correlations were determined for each task. Stepwise multiple linear regression, after controlling for the effects of the speech functions tested, revealed statistically significant influences of topical choice, novelty, and video animation on these measurement qualities in the sample of tasks in the current study. For example, topical choice had a slightly negative effect on task discrimination. On average, the presence of topical choice lowered the discrimination index by approximately 0.15, affecting Tasks 1, 7, and 11. This suggests that, even though the presence of topical choice may improve authenticity, it may have a slightly negative effect on measurement quality. The decrease could be related to the increase in the entropy of responses associated with having choice, inviting a greater range of topics and approaches to responding. However, novelty and, to a lesser extent, video animation both had slightly positive effects on measurement quality. The point–measure correlations of Tasks 4, 5, and 8 increased by an average of 0.04 with the presence of novelty, and video animation appeared to raise this correlation by another 0.01 in Task 5. Vocabulary support, on the other hand, had no detectable effects on measurement quality except for its previously mentioned contribution to task difficulty.

Table 4.9

Effects of Task Characteristics on Qualities of Measurement

Task Characteristic	Task Difficulty	Discrimination Index	Point–Measure Correlation
Topical choice	–0.37	–0.15	
Vocabulary support	+0.65		
Novelty	+0.99		+0.04
Video animation	+0.22		+0.01

Relationships Between Task Characteristics, Engagement, and Performance

While Stage 3 revealed the net effects of task characteristics on performance, a multiple-indicators multiple-causes (MIMIC) structural model was useful for exploring whether those effects were direct or whether they were mediated by engagement. After an initial model of task performance was specified, competing models including a task engagement factor were compared. With the final model, the direct and indirect effects of topical choice, vocabulary support, novelty, and video animation on performance were disentangled. Lastly, the model was tested for invariance across age groups. Consequently, Stage 4 expanded upon the findings of Stage 3 by providing further evidence about how performance was systematically influenced by task characteristics in a manner consistent with models of language performance and engagement.

The Initial Model

Since the endogenous variables (e.g., performance scores) failed to demonstrate multivariate normality (cf. DeCarlo, 1997), structural models were specified using maximum likelihood estimation with robust estimators. The first model specified was a model of task performance (see Figure 4.6). In this model, task performance was indicated by the two assigned ratings (scores), which were also influenced by the severity of the raters (previously measured by LLTM) and the number of points available according to the rubric type (i.e., 0-to-3 points or 0-to-5 points). Task performance itself was influenced by examinee ability and the task type

difficulty (also measured by LLTM). As expected, much of the variance in task performance (74.5%) was explained by examinee ability and task type difficulty alone.

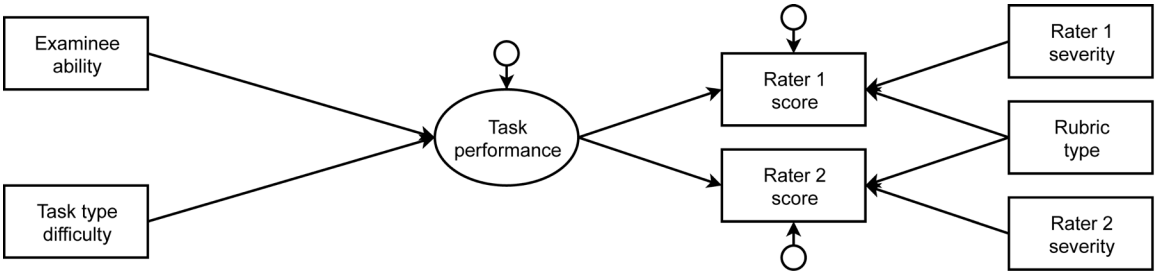


Figure 4.6. Initial model of task performance.

Competing Models

Next, the initial model was expanded to include task engagement. Task engagement was indicated by three acoustic features believed to likely signal engagement (i.e., the mean harmonicity, the mean shimmer, and the standard deviation of shimmer), and it was influenced by the four task characteristics hypothesized to support engagement (i.e., topical choice, vocabulary support, novelty, and video animation). Baseline, hypothesized, and alternate models were thus specified. The baseline model showed both task engagement and performance being influenced by the added task characteristics, but no relationship between task engagement and performance was specified (see Figure 3.3). The hypothesized model was identical to the baseline model except that task engagement was predicted to contribute to performance (see Figure 3.4). The alternate models were also identical to the baseline model, but the relationships between task performance and engagement were different from the hypothetical model (see Figures 3.5 and 3.6). The resulting fit statistics are presented in Table 4.10. Of the four models, the hypothesized model and the second alternate model had the highest CFI and TLI values and the lowest RMSEA and SRMR values, making them the best fitting models to the data. A chi-square difference test, using a scale correction factor as suggested by Sattora and Bentler (2010), revealed that model fit was not negatively affected by removing the path from task performance to engagement (i.e., the

path that distinguished the second alternate model from the hypothesized model), so the second alternate model was rejected in favor of the hypothesized model. Therefore, the hypothesized model, featuring a path drawn from task engagement to performance, suggests that not only was there a correlational relationship between task engagement and performance, but also task engagement likely supported performance in the current study.

Table 4.10

Comparison of Model Fit Statistics

Model	X^2	*df*	*p*	RMSEA	CFI	TLI	SRMR
Baseline	602.98	38	<0.001	0.06	0.97	0.96	0.03
Hypothesized	375.58	37	<0.001	0.05	0.98	0.97	0.02
Alternate 1	546.57	37	<0.001	0.06	0.97	0.96	0.02
Alternate 2	365.71	36	<0.001	0.05	0.98	0.97	0.02
Final	377.07	39	<0.001	0.04	0.98	0.98	0.02

All the paths in the hypothesized model were statistically significant except for the paths between vocabulary support and task engagement and between video animation and task engagement. To test whether these two nonsignificant paths should be included in the final model, the hypothesized model was re-specified with the paths removed one at a time. Chi-square difference tests, using a scale correction factor as before, revealed that model fit was not negatively affected by the removal of the paths. This suggests that while topical choice, vocabulary support, novelty, and video animation all had a direct effect on task performance, only topical choice and novelty—not vocabulary support or video animation—significantly affected task engagement, as indicated by acoustic measures of harmonicity and shimmer. The final model, complete with the standardized parameter estimates, is presented in Figure 4.7. See Appendix F for a list of the unstandardized coefficients.

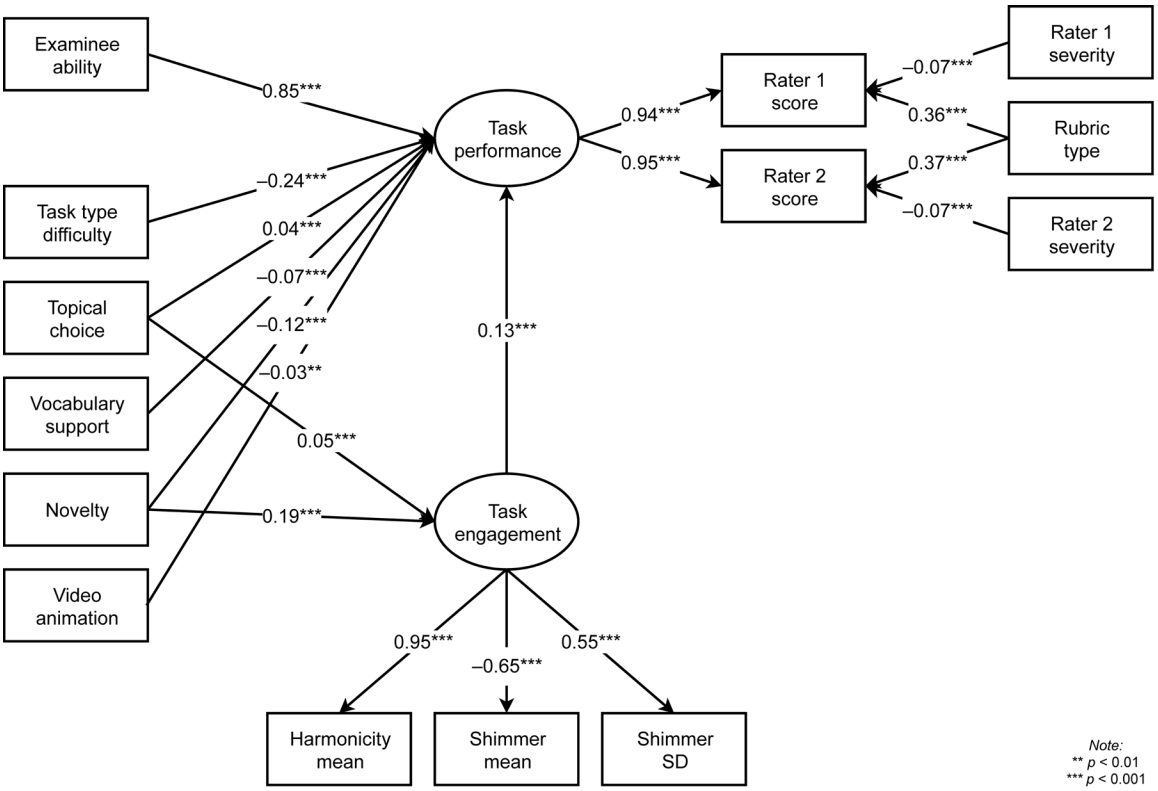


Figure 4.7. Final MIMIC model of task performance and engagement.

The final model, through its inclusion of a task engagement factor, reveals a more complete picture of the relationships between task characteristics and task performance than what was identified by LLTM in Stage 3. The inclusion of task engagement and the task characteristics of choice, vocabulary support, novelty, and video animation explained an additional 3.3% of the overall variance in performance compared to the initial model. It should be noted, however, that this relatively small change accounts for the effects of only the four task characteristics initially hypothesized to support engagement. Other task characteristics, such as the test being administered on a computer, likely also contributed to task engagement and performance, though this was not a focus of the current study. Thus, task engagement should be considered an important and relevant component in models of performance.

The Decomposition of Effects

The final MIMIC model reveals how some task characteristics may influence task performance and engagement differently. Although all four task characteristics being investigated had correlational relationships with performance, only choice and novelty demonstrated correlational relationships with engagement. Furthermore, while novelty seems to have had a positive effect on task engagement, its apparent effect on performance was negative. In order to further explore this difference, the total net effect of each task characteristic on performance was decomposed into its indirect effect (i.e., through its effect on engagement) and its direct effect.

Table 4.11 shows the standardized coefficients for the direct, indirect, and net effects of each task characteristic on task performance. Topical choice appears to have supported both task performance and engagement, with nearly one fifth of the net effect on performance stemming from the task characteristic's effect on engagement. Novelty, on the other hand, appears to have had a negative direct effect and a positive indirect effect on performance, meaning the overall net effect of topical choice on performance was negative at –0.10. This indicates that although some task characteristics may support task engagement, the net effects on overall performance can still be negative.

Table 4.11

Direct and Indirect Effects of Task Characteristics on Speaking Task Performance

Task Characteristic	Direct Effect	Indirect Effect	Net Effect
Topical choice	0.04	0.01	0.05
Vocabulary support	–0.07		–0.07
Novelty	–0.12	0.02	–0.10
Video animation	–0.03		–0.03

Multigroup Analyses

To explore whether the relationships among task characteristics, task performance, and task engagement were consistent for different examinees, model invariance across YLL age

groups was investigated. Five progressively stricter hypotheses of invariance were tested: First, a configural model was specified to confirm that the final model fit both the oldest and youngest age groups. Second, a metric model was specified to confirm that factor loadings were consistent between groups. Third, a scalar model was specified to confirm that the starting value of the scale the factor is based on was consistent between groups. Fourth, factorial invariance was tested to confirm that the overall error in the prediction of latent variables (e.g., task performance and engagement) was consistent between groups. Lastly, a residual error invariance test expanded upon the previous test by confirming that residual errors were consistent between groups. Chi-square difference tests (using scale correction factors as before) revealed how each model did not represent a significant degradation from the previous one. Results of the model invariance tests are presented in Table 4.12. Each of the five hypotheses was confirmed, indicating that the structure was remarkably stable across age groups. Such "strict invariance represents a highly constrained model and is rarely achieved in practice" (Bialosiewicz, Murphy, & Berry, 2013, p. 9), making claims for structural invariance across age groups even more compelling. In other words, the relationships among task characteristics, engagement, and performance were consistent between the youngest and oldest age groups in the current study. This suggests that any variation related to the range of ages targeted by the test had little or no effect on task performance or engagement.

Table 4.12

Results of Model Invariance Tests

Hypothesis	X^2 (*df*)	Scaled X^2 difference	RMSEA	CFI	TLI	SRMR
Configural invariance	283.21 (80)		0.04	0.98	0.97	0.02
Metric invariance	285.97 (83)	4.21	0.04	0.98	0.98	0.02
Scalar invariance	293.01 (88)	7.22	0.04	0.98	0.98	0.02
Factor invariance	305.93 (99)	11.98	0.04	0.98	0.98	0.02
Residual error invariance	311.57 (104)	8.19	0.04	0.98	0.98	0.02

Stage 4 teased apart the effects of topical choice, vocabulary support, novelty, and video animation on task performance from their effects on task engagement. While all of the task characteristics along with task engagement were shown to have effects on task performance, only topical choice and novelty were found to have any effect on task engagement, as indicated by likely acoustic indicators of engagement. Surprisingly, the relationships among these task characteristics, task performance, and task engagement were consistent across age groups. These findings serve as additional evidence to support the claim that task characteristics systematically contribute to performance in a manner consistent with the simplified model of speaking performance and engagement (see Figure 2.7). However, the results also indicate that task engagement should be incorporated into models of language proficiency because some task characteristics directly affect task performance and some task characteristics indirectly affect task performance (i.e., by directly affecting engagement, which in turn supports performance).

In YLLs' Own Words

To confirm and expand the findings about task engagement from Stage 4, a qualitative study was undertaken to investigate how YLLs actually experience tasks that contain topical choice, vocabulary support, novelty, and video animation. Retrospective structured interviews gave a small group of YLLs (n=8) a chance to describe their perceptions of tasks and task characteristics in their own words. A detailed overview of YLLs' self-reports is presented below. Italics are used to indicate language that has been translated into English.

In response to questions about their overall impression of the test, participants responded very favorably. The most commonly used word to describe the tasks was "fun." One participant stated that "*this is the most fun test I have ever taken.*" Another indicated that "*it didn't feel like a test.*" A third even said, "I wanted to do it again." The responses indicated that the computerized testing format was well-received by the YLLs.

However, when asked about favorite tasks, participants did not respond so uniformly. Task 1 (What's your favorite animal?), Task 3 (What's strange about the bus?), and Task 5 (What happened to the key?) were all repeatedly identified as tasks that were well-liked. Task 7 (Questions about the tiger) was mentioned as a least favorite task. Being able to make a choice about a favorite animal to discuss was commonly cited as a reason for liking Task 1. Tasks 3 and 5 were frequently described as "funny." One participant alluded to the game-like quality of Task 3 by comparing it to Golden Bell, a popular quiz show in Korea. However, another complained that the strange things examinees were supposed to report were "*so obvious that it wasn't very interesting.*" The video animation in Task 5 was liked by one participant because it was "*a fiction… If you tug on a flower, there is no way it wouldn't get pulled out of the ground.*" Another participant was visibly excited when describing the novel elements of the video animation. "It's fun because Billy is so genius. Billy know that when he pull the branch, the ladder comes out. I don't know why… inside the tree there's a ladder." Finally, reasons that participants did not like Task 7 included "The tiger didn't have any color" and "I'm not very interest of the tiger… *I would like to speak about the animal that I like the most.*" One participant suggested that the task could be improved if he "*could get answers to the questions [he asked]*."

Every YLL interviewed reported liking tasks that featured topical choice. One participant described how "*if there were only one choice, it wouldn't have been fun.*" Another participant explained, "*Normally, the teacher decides what you should talk about, but in this case, since I could choose what I wanted to talk about, it was more fun.*" Reports like these confirm the finding from Stage 4 that topical choice supports both agency and engagement. However, one participant revealed that "*if there were an option to choose wolf, I would have… I like wolves better than tigers, but since there were no wolves I chose the tiger.*" Another one indicated that "*if there are more options than three, it would make it more interesting.*" The effect of varying the number of options on YLLs' engagement is not clear, though nobody mentioned the task that was

open-ended, suggesting that some constraints on the number of options may be helpful for children in this age range.

Vocabulary support, on the other hand, was not so well received. Some participants recognized how the introduction of relevant vocabulary could benefit their performance. One participant mentioned how "*it helps when you're speaking… If there is a word that I don't understand, that would make me upset.*" Another participant indicated how "*because there was an explanation, I could use them [the words]. If you use the words well, your sentences improve.*" However, when asked how vocabulary support affected participants' enjoyment of the task, responses were not enthusiastic, ranging from "It's okay" to "not that fun" to "*it was not fun because I like getting to the questions right away.*" The negative view of this task characteristic mirrors the lack of a relationship between vocabulary support and task engagement found in Stage 4.

Novelty was the task characteristic discussed most by YLLs in the study. One participant said, "*I think it was fun because it didn't make sense.*" Another participant explained how "*it helps our creativity, and it makes us feel good.*" A third participant related how "*it's a little weird... It would never happen in real life, but it's in the test so it's fascinating.*" Just as these comments indicate, novelty was identified as having a positive effect on task engagement in Stage 4. However, novelty also had a negative effect on task performance. Some YLLs in the study made the same observation. For example, one participant remarked how "*it was hard to explain [everything],*" and another commented on his struggle "*to explain a lot… to explain exactly what happened.*" These reflections on performance correspond to the negative effect of novelty on performance identified in Stages 3 and 4.

Lastly, the presence of video animation was described by different participants as making the task "fun" and "*easy to understand*" and "*more real*." "*In most tests, we don't have videos, but this test is far more interesting because there are videos*," explained one participant. Another

participant even suggested that "*we need a 3-D video*." In contrast to Stage 4, which did not reveal a relationship between video animation and task engagement, the qualitative phase suggested that video content was actually liked by participants. Why this mismatch? One possibility is that, because the test was computer administered, the difference between tasks with successive static images and those with more dynamic video imagery may not have been very pronounced. Another possibility is that video content merely served as "eye candy" (Trushell, Burrell, & Maitland, 2001), making the content likable but with little actual impact on task engagement. But mostly, participants seemed more attentive to the novel elements contained within the video animation than in the video animation itself. Evidence for this last explanation exists in the fact that none of the participants mentioned the presence of the video until explicitly asked about it toward the end of the interviews. However, since only one task contained a video animation, it would be reckless to draw any firm conclusions except that this topic warrants further investigation.

Summary

The current study provided evidence regarding the validity of the *TOEFL Primary* Speaking test and other assessments like it. Specifically, evidence was gathered about the evaluation, generalizability, and explanation inferences in a validity argument. In addition, the influences of the task characteristics of topical choice, vocabulary support, novelty, and video animation on task engagement were investigated both quantitatively and qualitatively. The results help fill some gaps in the literature regarding assessing young learner speaking ability and the nature of task engagement, but they also point to the need for additional research. These are discussed more fully in Chapter 5.

Chapter V

DISCUSSION AND CONCLUSION

The purpose of the current study was to investigate examinee age, task characteristics, examinee engagement with task content, and speaking performance in the context of a technology-enhanced speaking test for young learners. In the process, evidence related to claims about the validity of the *TOEFL Primary* Speaking test was gathered, and factors and products of engagement were explored. In this chapter, key findings are summarized. Then, theoretical, methodological, and practical implications of the study are discussed. Lastly, limitations of the study and suggestions for further research are described.

Key Findings

Building upon previous work defining the language use domain among young learners of English as a foreign language (e.g., Cho et al., 2016, 2017; Turkan & Adler, 2011), evidence related to the evaluation, generalization, and explanation inferences of a test validity argument was collected in the course of addressing the first three research questions, respectively. Therefore, key findings are presented along with how they back assumptions that underlie claims about performance evaluations, the generalizability of scores, and an explanation into the meaning of scores following models set by Chapelle, Enright, and Jamieson (2010) and Knoch and Chapelle (2018). The last two research questions involved taking a closer look at the nature of engagement—how it may be affected by certain task characteristics, how it may affect likely acoustic indicators of engagement, and how it may relate to examinee age and performance. Findings related to engagement can be viewed as further contributing to an explanation into the meaning of scores.

The first research question about how well technology-enhanced speaking test tasks discriminate among YLLs according to their speaking abilities is directly related to claims about the validity of performance evaluations. Several assumptions underlie the claim that the evaluation of performances results in scores with the intended characteristics. Table 5.1 lists some of these assumptions on the left, and relevant evidence gathered in the current study supporting each assumption is listed on the right. For example, Rasch analysis revealed a high person-separation index, serving as evidence that the test effectively discriminated among examinees of varying ability. In addition, bias analyses revealed no major interactions between examinee age and tasks or examinee age and raters, indicating that the resulting task scores were appropriate regardless of examinee age. These findings from many-facet Rasch analyses effectively addressed the first research question.

Table 5.1

Assumptions and Backing Related to the Evaluation Inference

Assumptions	Backing
Tasks, raters, and rating scales effectively discriminate among examinees of varying ability.	Many-facet Rasch measurement reveals adequate person separation on the scale criteria.
Tasks demonstrate psychometric unidimensionality.	Tasks reflect good model fit.
Raters can consistently apply rating scales to responses.	Fit statistics indicate a high degree of rater consistency.
Rating scales are appropriate for assigning scores to responses.	Rating scale boundaries are well spaced and monotonic.
Tasks function consistently between different examinee groups.	Bias analyses reveal that tasks function similarly for older and younger examinees.
Raters can consistently apply rating scales to different examinee groups.	Bias analyses show raters applying rating scales consistently across age groups.

The second research question about the consistency and dependability of scores is related to the generalizability inference. Findings from a series of generalizability and decision studies back assumptions that underlie the claim that scores are estimates of expected scores across tasks, ratings, and forms. Table 5.2 lists some assumptions and related evidence collected in the current study. For example, the percentage of score variance attributable to examinees was substantially greater than that attributable to the tasks, ratings, or myriad interactions with tasks and ratings. Furthermore, adding an additional task scored on a scale of 0 to 5 appeared to maximize score dependability without unduly lengthening the test, mirroring the actual configuration of operational *TOEFL Primary* Speaking test forms.

Table 5.2

Assumptions and Backing Related to the Generalizability Inference

Assumptions	Backing
Assigned scores are consistent across parallel tasks and ratings.	Results from a generalizability study indicate that the bulk of score variance comes from examinees, not from tasks or ratings.
The configuration of tasks and ratings provide stable estimates of examinee performances across test forms.	Decision studies indicate that operational test forms contain optimal numbers of tasks and ratings for maximizing score dependability.
The relative contributions of different task types to total score variance do not vary according to age.	The effective weights of the different task types on total score variance are consistent across age groups.

The third research question asks how certain task characteristics hypothesized to support engagement affect measurement qualities like difficulty. This question directly relates to the explanation inference of a test validity argument. This inference is based on a claim about how expected scores are attributable to the underlying construct. Assumptions and backing for this claim are presented in Table 5.3. Since contemporary models of speaking performance predict how task characteristics interact with personal attributes to influence performance, the task

characteristics of topical choice, vocabulary support, novelty, and video animation should systematically affect performance. Results from Fischer's (1973, 1995) linear logistic test model (LLTM) indicate that, in the current study, topical choice systematically made tasks easier, but the other three task characteristics systematically made them more difficult. In addition, scores on tasks with novelty and video animation tended to correlate a bit more strongly with ability estimates than scores on tasks without novelty or animation did, suggesting that tasks with novelty or video animation were slightly more sensitive to the underlying construct.

Table 5.3

Assumptions and Backing Related to the Explanation Inference

Assumptions	Backing
Task difficulty is systematically influenced by task characteristics.	LLTM shows how speech functions and other task characteristics systematically influence task performance.
The abilities of tasks to elicit evidence of the intended construct can be systematically influenced by task characteristics.	The presence of novelty or video animation in tasks results in stronger point-measure correlations.
Age has little effect on expected scores.	Structural invariance across age groups suggests that performance is not unduly affected by examinee age.

The last two research questions called for quantitative and qualitative investigations related to the role of engagement in an explanation into the meaning of scores. Key findings from these two investigations are presented in Table 5.4. First, potential indicators of engagement were extrapolated from comparing 100 acoustic features from spoken responses to scores on the post-test engagement survey. Three likely indicators of engagement were retained and used to create a structural model to disentangle the effects of certain task characteristics on engagement from those on performance. A multiple-indicators multiple-causes (MIMIC) model revealed how only topical choice and novelty appeared to support engagement in the current study—findings

that were corroborated qualitatively through retrospective interviews with YLLs. This corroboration also serves as evidence for the validity of the three likely indicators of engagement used in the model. The model also shows how engagement did appear to support performance, and, surprisingly, all of these relationships were found to be invariant across age groups, suggesting that both engagement and performance were not heavily influenced by age.

Table 5.4

Findings Related to Engagement

Model components	Findings
Acoustic indicators of engagement	Relative increases in both the mean harmonicity and the standard deviation of shimmer along with a relative decrease in mean shimmer appear to signal engagement.
Task characteristics	Structural modeling suggests that topical choice and novelty support engagement, but vocabulary support and video animation do not. These findings were also reflected in retrospective verbal reports by YLLs.
Performance	Structural modeling reveals that engagement likely supports performance.
Age	Structural invariance across age groups suggests that engagement was not affected by examinee age.

In summary, the findings related to evaluation, generalization, and explanation claims of a validity argument for a technology-enhanced speaking test for YLLs largely mirror findings from other studies involving adult learners. For example, *TOEFL Primary* Speaking test tasks demonstrated good fit to the Rasch model just as speaking tasks on the *TOEFL* iBT have (e.g., Winke et al., 2011, 2013). In both cases, tasks and raters functioned as expected, though confirming the lack of any serious interactions with examinee age is a new finding of the current

study. Furthermore, increasing the number of tasks in place of adding more ratings of responses to the same tasks was shown to maximize the dependability of speaking scores on the *TOEFL* test (Lee, 2005, 2006), a finding mirrored by the current study. However, when it comes to explanation claims, most previous research with adults investigated the relationships between speaking and the other domains of reading, listening, and writing (e.g., Sawaki, Stricker, & Oranje, 2008, 2009; Stricker & Rock, 2008). Though some efforts have been made to examine how characteristics of constructed-response prompts affect task difficulty (e.g., Cho, Rijmen, & Novák, 2013), the current study was unique in that it also examined how they affect examinee engagement with tasks. At no point was any evidence found rebutting the three validity claims investigated.

Implications

The current study provided support for evaluation, generalization, and explanation claims in a validity argument for the *TOEFL Primary* Speaking test. However, the study has other implications, too. For example, the findings point to a need to develop a theoretical model of task engagement and to consider how it relates to models of language performance. In addition, the study represents an advance in terms of engagement detection in the context of a speaking test for young learners. Furthermore, the study has some practical implications for test development. In this section, theoretical, methodological, and practical implications of the study are discussed.

Theoretical Implications

In Chapter 2, a taxonomy of four types of task characteristics that may support engagement was proposed. This taxonomy includes task characteristics that support a sense of agency, a sense of success, situational interest, and social connectedness. The current study undertook to investigate how examples of the first three types of task characteristics would influence engagement. In terms of supporting a sense of agency, the presence of topical choice in

test tasks in the current study likely had a positive effect on engagement. Novel elements stimulating situational interest in tasks likewise had a similar effect. However, vocabulary support did not seem to have any effect on engagement, which suggests either that vocabulary support, as operationalized on the *TOEFL Primary* Speaking test, may not really support a sense of success or that supporting a sense of success may not really support engagement. The former possibility is more likely; providing an opportunity to select vocabulary support only as needed might have yielded different results (see Poehner, Zhang, & Lu, 2017). Also, video animation by itself did not seem to support engagement as much as the novel elements contained within it did. This suggests that the content may be more important than the presentation when it comes to stimulating engagement. Of course, the proposed taxonomy of task characteristics that may support engagement warrants further study to confirm and expand upon the findings of the current study.

The taxonomy of task characteristics can also inform a theoretical model of task engagement. Figure 5.1 shows how task characteristics interact with learner motivation and personal interests to form a componential model of engagement. Although the current study revealed how task characteristics that support a sense of agency and situational interest likely contribute to task engagement, it is not yet clear whether task characteristics that support a sense of success or that support social connectedness do as well. Furthermore, though it was beyond the scope of the current study, the role of personal attributes, like learner motivation and personal interest, in task engagement could also be relevant to a research agenda. The current study does suggest, however, that age does not play a significant role in learner engagement. The model of task engagement described here reflects one of the only known attempts to illustrate which factors go into a learner's engagement with a task, but it may also have relevance to research in other noneducational contexts, such as workplace engagement or user engagement with information technologies.

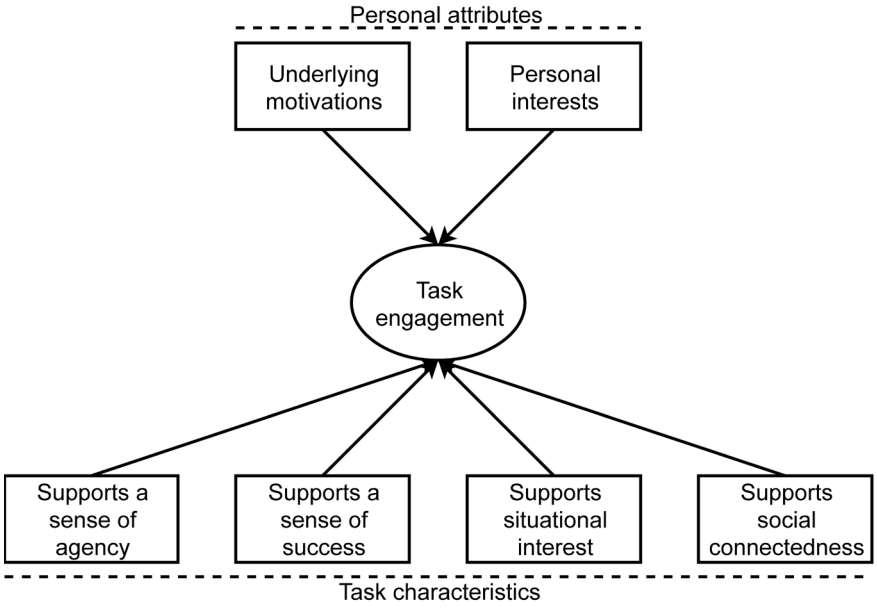


Figure 5.1. A model of task engagement.

Along with the need to fill out a model of task engagement is a need to understand how it fits into a model of language performance. The current study demonstrated how engagement and the task characteristics that support it likely play a role in young learner performance on a language test. Therefore, models of language performance should feature engagement as a separate but relevant component that influences performance. Some models of performance (e.g., Bachman, 1990; Bachman & Palmer, 1996, 2010; Turner & Purpura, 2016) already include affective variables, but it is not clear what contributes to them or how they affect performance other than through the mysterious process of mediation. However, engagement contains a cognitive dimension in addition to an affective one (Fredricks et al., 2004; Fredricks, Filsecker, & Lawson, 2016), and more conceptual work is needed to better understand the possible distinctions between affective and cognitive engagements. The roles that such affective and cognitive factors play in language performance would likely have implications for explanation claims of a test validity argument.

Methodological Implications

The current study also has some methodological implications. The primary example of this is in the use of acoustic cues from spoken responses themselves as signals of engagement. Also, unlike most research on acoustic indictors of emotion that relies on speech data from voice actors (Scherer, 2013), the investigations in the current study involved a corpus of naturally occurring YLL test responses in the quest for dependable indicators of task engagement. Out of 100 acoustic, prosodic, and temporal features considered as potential indicators of engagement (see Appendix D), the mean harmonicity, the mean shimmer, and the standard deviation of shimmer were significantly correlated with reported engagement at the whole test level. Although using transformations of these measures to identify engagement at the individual task level was rather inferential (i.e., an assumption was made that acoustic features that correlated with whole-test engagement would also signal individual-task engagement), the structural model built with them was highly predictive of YLL's qualitative reports about how engaging various tasks and their task characteristics were. This triangulation suggests that these measures may be worthy indicators of engagement, matching the findings of some previous studies (e.g., Kim & Truong, 2016; Pérez et al., 2016). These and possibly other features extracted from spoken responses themselves may offer deeper insights into YLLs' experiences of test content. However, these results call into question other previously used but not validated measures of engagement in the literature, like time on task (e.g., Lambert et al., 2017; Phung, 2017), which was not identified as a good predictor of engagement.

Practical Implications

In addition to the theoretical and methodological implications of the current study and its findings, some practical implications also exist. These can roughly be divided into three categories: implications related to test validity, implications for test development, and implications beyond assessment. Each of these is discussed here in turn.

Implications related to test validity. The principal finding related to test validity is that examinee age had little influence on or interaction with the performance of tasks and raters on *TOEFL Primary* Speaking test tasks. In addition, the influence of task characteristics on task engagement and performance was remarkably stable across examinee age groupings. These findings indicate that, contrary to concerns that helped motivate the study and despite age-related developmental variations across the sample of examinees, the *TOEFL Primary* Speaking test tasks were appropriate for measuring the language abilities of YLLs across a wide age range. Thus, validity evidence gathered in support of evaluation, generalization, and explanation claims about test usefulness is not threatened by the fact that the test targets YLLs between the approximate ages of 8 and 13 years.

Implications for test development. The findings also have implications for assessment design. In particular, test developers may find this and other research on task characteristics to be particularly informative when creating test and task specifications. Eventually, a validated taxonomy of task characteristics that support engagement could help test and other content developers employ a more data-driven approach to selecting task characteristics in light of their effects on task engagement in addition to the linguistic demands of the task. The current study investigated four task characteristics: topical choice, vocabulary support, novelty, and video animation. Here are some comments related to each one.

Topical choice. In the current study, only topical choice was shown to have positive effects on both engagement and performance. However, the presence of topical choice came with some tradeoffs in terms of task discrimination. While topical choice did seem to make tasks easier for examinees, it, unsurprisingly, introduced some variability that may account for a decrease in discrimination. Giving options to examinees on a test does appear to support engagement, suggesting that other task characteristics that support a sense of agency may also be desirable when developing technology-enhanced speaking test tasks and potentially other kinds of tasks. Of

course, the use of choice should be judicious as evidence suggests that posing too much choice can be detrimental to performance (e.g., Iyengar & Lepper, 2000; Mozgalina, 2015). The effects of varying the number and types of choices on engagement and performance warrant additional study.

Vocabulary support. Vocabulary support, on the other hand, did not have positive effects on task engagement or performance in the current study. There was no evidence that it improved measurement quality, either. In fact, vocabulary support seems to have functioned more as additional stimulus material than anything else, and it could potentially disadvantage younger students if it increased the cognitive load. For these reasons, the use of vocabulary support in speaking assessment tasks for young learners should be judicious and limited or be implemented in a different way (e.g., by making the vocabulary support available on an as-needed basis). More research is needed to see whether other task characteristics intended to support examinee success have any effect on engagement and performance.

Novelty. Like topical choice, novelty also seemed to have had a positive effect on engagement in the current study. Novelty was also correlated with increased point-measure correlations on the sample of tasks. However, novelty actually had the strongest negative influence of the four task characteristics investigated on performance. This finding suggests that some task characteristics can affect task engagement and performance differently. Careful consideration may be required to confirm that its use, including the resulting increase in the linguistic demands of a task from obligating novel language production, is construct relevant. More investigations are needed to see if such findings generalize to YLL contexts beyond the current study.

Video animation. Lastly, video animation appeared to have little or no effect on task engagement. However, it did seem to improve the point-measure correlation slightly. Also, its presence did not interact with examinee age, which suggests that any concerns about the possible

memory load that a dynamic video, unlike a static image, might have on younger examinees taking the current test may not be justified. However, with only one task that featured this task characteristic, it is difficult to draw any firm conclusions. Again, more study is needed.

Implications beyond assessment. The automatic detection of user engagement could also have many uses beyond the realm of speaking assessment tasks. For example, the findings will also likely generalize to pedagogical and technological design in other contexts, such as computer-assisted language learning, interactive video game design, and even voice assistants. Such affect recognition capabilities could potentially support more authentic and responsive experiences for users. Some work has already begun with recognizing common emotions like joy and anger (e.g., Schuller, Steidl, & Batliner, 2009; Schuller et al., 2010), but the current study is one of the first to suggest the possibility of automatically measuring user engagement with specific tasks by using acoustic indicators from naturally occurring spoken responses from young learners. Acoustic indicators of engagement could be yet another metric to consider when developing responsive and adaptive instructional content.

Limitations and Directions for Future Research

The current study was not without limitations. The most prominent ones are described below. The limitations can generally be categorized as relating to the participants, the instruments, the procedures, or the analyses. Each category is discussed in turn along with some suggestions for further research.

In terms of the participants, one significant limitation was the rather narrow sampling of YLLs for the qualitative phase of the investigation. More specifically, although the test tasks were intended for young learners between the ages of approximately 8 and 13 years and from a variety of international contexts, the sample of participants were all around 8 or 9 years old, and they all hailed from the same country (i.e., South Korea). Although their feedback did largely agree with

the quantitative findings, further study involving some older children and other national backgrounds could serve to strengthen confidence in the results.

In addition, the samples for both the quantitative and qualitative phases of the investigation involved young learners in several English-as-a-foreign-language contexts, but there is also interest in developing high-quality technology-enhanced assessments in English-as-a-*second*-language contexts. However, in terms of personal attributes, the current study was primarily focused on effects related to young learner age. Given the limited life experiences of YLLs, diverse cultural, educational, and digital literacy backgrounds of examinees are also likely relevant to test performance, levels of engagement, and possibly other cognitive and affective states, such as anxiety. Investigations into these factors are still needed in order to generate evidence that supports or rebuts assumptions underlying claims for test validity.

In terms of the instruments, several limitations were artifacts of the ex post facto research design. For example, the number of test tasks was small, so the range of task characteristics that was explored was also small, thus the generalizability of findings may be limited. For example, the effects of varying the scenario around which tasks are built are still unknown. Furthermore, experimentally manipulating task characteristics in the tasks was not possible, which also limited possible investigations into the effects of task characteristics and the robustness of any conclusions drawn. In addition, many more task characteristics could potentially be examined with more tasks, especially if the tasks are designed with the proposed taxonomy of task characteristics that may support engagement in mind. Future studies would do well to deliberately develop speaking tasks to more fully investigate the effects of certain task characteristics on engagement and performance among YLLs.

Even more limiting was the fact that the post-test engagement survey consisted of only two dichotomous items that covered the whole test. Based on this survey, acoustic indicators of task engagement could only be inferred. As a result, the connections between the acoustic

features and the levels of engagement they are thought to reflect were not particularly strong. Although retrospective interviews with a small set of YLLs supported the quantitative findings, more detailed information about examinees' engagement with test content could bolster claims about how acoustic and other features of responses relate to engagement. For example, the post-test engagement survey could have been improved if it consisted of Likert-type questions, which would not have been much more obtrusive than the dichotomous ones presented. In addition, if the data had been collected immediately after and about each task instead of at the end of the test, the potential to identify more sensitive measures of engagement might have existed. Of course, increasing the number of survey items would also allow for a greater range of child-centered terms to be used in place of the word "engaging" besides just "fun" and "like." Since engagement is commonly thought of as having multiple dimensions (Fredricks et al., 2004; Svalberg, 2009), a richer variety of survey items could potentially tap into cognitive dimensions of engagement in addition to affective ones. Much work remains to be done in terms of identifying objective measures of task engagement. Going forward, task-level feedback and acoustic features could also be coupled with an expanded array of acoustic, eye-tracking, electroencephalogram, functional magnetic resonance imaging, and other biometric data as potential signals of engagement.

Finally, in terms of limitations related to the analyses, traditional structural equation modeling was used based on three likely acoustic indicators of engagement. However, using machine learning as an analytic tool could potentially expand the number of acoustic and other features considered to make even more accurate classification decisions. The use of more sophisticated data-mining techniques, such as support vector machines, has already given rise to a whole field known as affective computing (e.g., Eyben, 2016). Automated approaches for detecting emotion could very naturally cross-pollinate with engagement research. Additional

studies would do well to take advantage of the affordances of machine learning as an analytic tool when examining the nature of engagement on tests for young learners.

Summary

In conclusion, three principal findings of the current study are relevant to the continued growth and expansion of teaching and assessment for the population of YLLs. First, age had surprisingly little effect on test task performance, rater behavior, and engagement. This suggests that the window for what can be considered developmentally appropriate content for language tasks may be a bit bigger than once thought. Second, of the task characteristics studied, only choice seemed to improve levels of both engagement and performance, while novelty appeared to improve levels of engagement only. Data generally support the use of these task characteristics in other assessments going forward. Lastly, acoustic measures related to harmonicity and shimmer seemed to function well as indicators of engagement. These features, extracted from spoken responses themselves, may offer deeper insights into YLLs' experiences of test content.

Continued research on YLL assessments is especially important because of the large numbers of YLLs taking these assessments every day and the adults making decisions about and as a result of these assessments. Therefore, in addition to confirming the findings and addressing the limitations of the current study, evidence to support extrapolation and utilization inferences is needed in order to complete an argument for the test's validity. Extrapolation claims relate the measured constructs to the target language use domain. This could be accomplished by matching test performance with teacher observations of classroom performance or even with performance on other validated assessments. Utilization claims indicate that decisions are appropriate and beneficial for stakeholders. As technology-enhanced testing continues to grow in prevalence and sophistication, research into its validity and the role of engagement in such assessments must keep pace.

REFERENCES

Acha, J. (2009). The effectiveness of multimedia programmes in children's vocabulary learning. *British Journal of Educational Technology*, *40*(1), 23–31. https://doi.org/10.1111/j.1467-8535.2007.00800.x

Ainley, M. (2012). Students' interest and engagement in classroom activities. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 283–302). https://doi.org/10.1007/978-1-4614-2018-7_13

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology*, *5*(1), 202–232.

Anderson, V. A., & Lajoie, G. (1996). Development of memory and learning skills in school-aged children: A neuropsychological perspective. *Applied Neuropsychology*, *3*(3–4), 128–139. https://doi.org/10.1207/s15324826an0303&4_5

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/bf02293814

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, *44*(5), 427–445. https://doi.org/10.1016/j.jsp.2006.04.002

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. New York, NY: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct-validation of some components of communicative proficiency. *TESOL Quarterly*, *16*(4), 449–465. https://doi.org/10.2307/3586464

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/s1364-6613(00)01538-2

Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208. https://doi.org/10.1016/s0021-9924(03)00019-4

Baddeley, A. D. (2015). Working memory in second language learning. In Z. Wen, M. B. Mota, & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (pp. 17–28). https://doi.org/10.21832/9781783093595

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The Psychology of Learning and Motivation*, *8*, 47–89. https://doi.org/10.1016/s0079-7421(08)60452-1

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences*, *43*, 100–105. https://doi.org/10.1016/j.lindif.2015.09.001

Bagwell, C. (2015). SOund eXchange (Version 14.4.2). Retrieved from http://sox.sourceforge.net

Baïdak, N., Borodankova, O., Kocanova, D., & Motiejunaite, A. (2012). *Key data on teaching languages at school in Europe*. https://doi.org/10.2797/83967

Bailey, A. L. (2008). Assessing the language of young learners. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education* (2nd ed., Vol. 7, pp. 2509–2528). https://doi.org/10.1007/978-0-387-30424-3_188

Bailey, A. L. (2017). Theoretical and developmental issues to consider in the assessment of young learners' English language proficiency. In M. K. Wolf & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners* (pp. 25–40). https://doi.org/10.4324/9781315674391-2

Ballard, L., & Lee, S. (2015, March). *How young learners respond to computerized reading and speaking tasks*. Presented at the Language Testing Research Colloquium, Toronto, ON.

Banerjee, H.-T. L. (2019). *Investigating the construct of topical knowledge in a scenario-based assessment designed to simulate real-life second language use* (Doctoral dissertation). Teachers College, Columbia University, New York, NY.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path models of vocal emotion communication. *PLoS One*, *10*(9). https://doi.org/10.1371/journal.pone.0136675

Bergin, D. A. (1999). Influences on classroom interest. *Educational Psychologist*, *34*(2), 87–98. https://doi.org/10.1207/s15326985ep3402_2

Bialosiewicz, S., Murphy, K., & Berry, T. (2013, October). *Do our measures measure up? The critical role of measurement invariance*. Presented at the American Evaluation Association, Washington, DC.

Bialystok, E. (2002). On the reliability of robustness. *Studies in Second Language Acquisition*, *24*(03), 481–488. https://doi.org/10.1017/s0272263102003054

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bishop, R. S. (1990). Mirrors, windows, and sliding glass doors. *Perspectives*, *6*(3), ix–xi.

Blanco, J., & Howden, D. (2011). Seeking stakeholders' views on Cambridge English exams: School sector. *Cambridge ESOL: Research Notes*, *46*, 7–9.

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer (Version 6.0.13). Retrieved from www.praat.org

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge, Taylor and Francis Group.

Bongaerts, T., Van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*(4), 447–465. https://doi.org/10.1017/s0272263197004026

Brennan, R. L. (2001a). *Generalizability theory*. New York, NY: Springer.

Brennan, R. L. (2001b). MGENOVA (Version 2.1). Iowa City, IA: The University of Iowa.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341–366. https://doi.org/10.1177/0265532209104666

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research & Evaluation*, *22*(1), 1–10.

Butler, Y. G. (2009). How do teachers observe and evaluate elementary school students' foreign language performance? A case study from South Korea. *TESOL Quarterly*, *43*(3), 417–444. https://doi.org/10.1002/j.1545-7249.2009.tb00243.x

Butler, Y. G. (2016). Assessing young learners. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment* (pp. 359–375). Boston, MA: De Gruyter Mouton.

Butler, Y. G. (2017a). Challenges and future directions for young learners' English language assessments and validity research. In M. K. Wolf & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners* (pp. 255–273). https://doi.org/10.4324/9781315674391-14

Butler, Y. G. (2017b). Motivational elements of digital instructional games: A study of young L2 learners' game designs. *Language Teaching Research*, *21*(6), 735–750. https://doi.org/10.1177/1362168816683560

Bygate, M., Skehan, P., & Swain, M. (Eds.). (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. Harlow, UK: Routledge.

Cambridge ESOL. (2007). *Cambridge Young Learners English tests: Handbook for teachers*. Retrieved from http://www.cambridgeenglish.org/exams-and-qualifications/young-learners

Cameron, L. (2001). *Teaching languages to young learners*. Cambridge, UK: Cambridge University Press.

Cameron, L. (2003). Challenges for ELT from the expansion in teaching children. *ELT Journal*, *57*(2), 105–112. https://doi.org/10.1093/elt/57.2.105

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., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/i.1.1

Carreira, J. M. (2006). Motivation for learning English as a foreign language in Japanese elementary schools. *JALT Journal*, *28*(2), 135–157.

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English Proficiency of Foreign Students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language Testing Symposium: A Psycholinguistic Perspective* (pp. 47–69). London, UK: Oxford University Press.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & McOwan, P. W. (2010). Affect recognition for interactive companions: Challenges and design in real world scenarios. *Journal on Multimodal User Interfaces*, *3*(1–2), 89–98. https://doi.org/10.1007/s12193-009-0033-5

Chan, D. Y. C., & Wu, G. C. (2004). A study of foreign language anxiety of EFL elementary school students in Taipei County. *Journal of National Taipei Teachers College*, *17*(2), 287–320.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces Between Second Language Acquisition and Language Testing Research* (pp. 32–70). https://doi.org/10.1017/CBO9781139524711.004

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chaspari, T., & Lehman, J. F. (2016, September). *An acoustic analysis of child-child and child-robot interactions for understanding engagement during speech-controlled computer games*. Presented at the Interspeech Conference, San Francisco, CA. https://doi.org/10.21437/interspeech.2016-85

Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing*, *21*(3), 360–389. https://doi.org/10.1191/0265532204lt288oa

Cho, Y., & Getman, E. P. (2013, September). *TOEFL Primary Speaking: Use of technology to support scenario-based assessment activities for young EFL learners*. Presented at the TOEFL Research Symposium, Educational Testing Service, Princeton, NJ.

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2016). *Designing the TOEFL Primary Tests* (Research Memorandum No. RM-16-02). Princeton, NJ: Educational Testing Service.

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2017). Designing the TOEFL Primary tests. In M. K. Wolf & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners* (pp. 41–58). https://doi.org/10.4324/9781315674391-3

Cho, Y., Rijmen, F., & Novák, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT integrated writing tasks. *Language Testing*, *30*(4), 513–534. https://doi.org/10.1177/0265532213478796

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, *80*(2), 183–198. https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

City & Guilds. (2008). *ESOL for Young Learners: Qualification handbook*. Retrieved from http://www.cityandguilds.com

Clark, B. A. (2000). First- and second-language acquisition in early childhood. In D. Rothenberg (Ed.), *Issues in Early Childhood Education: Curriculum, Teacher Education, & Dissemination of Information* (pp. 181–188). Champaign, IL: University of Illinois at Urbana-Champaign.

Clark, J. L. D. (1979). Direct vs semi-direct tests of speaking ability. In F. B. Hinofotis & E. J. Brière (Eds.), *Concepts in Language Testing: Some Recent Studies.* (pp. 35–49). Washington, DC: TESOL.

Clément, R. (1986). Second language proficiency and acculturation: An investigation of the effects of language status and individual characteristics. *Journal of Language and Social Psychology*, *5*(4), 271–290. https://doi.org/10.1177/0261927x8600500403

Coplan, R. J., Prakash, K., O'Neil, K., & Armer, M. (2004). Do you "want" to play? Distinguishing between conflicted shyness and social disinterest in early childhood. *Developmental Psychology*, *40*(2), 244–258. https://doi.org/10.1037/0012-1649.40.2.244

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715–730. https://doi.org/10.1037/0022-0663.88.4.715

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley & Sons.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Crozier, W. R., & Hostettler, K. (2003). The influence of shyness on children's test performance. *British Journal of Educational Psychology*, *73*(3), 317–328. https://doi.org/10.1348/000709903322275858

Curtain, H., & Dahlberg, C. A. (2010). *Languages and children: Making the match* (4th ed.). Boston, MA: Pearson.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367–396. https://doi.org/10.1177/0265532209104667

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, *43*(3), 393–415. https://doi.org/10.1002/j.1545-7249.2009.tb00242.x

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*(3), 292–307. https://doi.org/10.1037/1082-989x.2.3.292

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. https://doi.org/10.3758/brm.41.2.385

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*(4), 499–533. https://doi.org/10.1017/s0272263100004022

D'Mello, S. K., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist*, *52*(2), 104–123. https://doi.org/10.1080/00461520.2017.1281747

D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, *20*(2), 147–187. https://doi.org/10.1007/s11257-010-9074-4

Dörnyei, Z. (2002). The motivational basis of language learning tasks. In P. Robinson (Ed.), *Individual Differences in Second Language Acquisition* (pp. 137–158). Amsterdam, The Netherlands: John Benjamins.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Dörnyei, Z., & Kormos, J. (2000). The role of individual and social variables in oral task performance. *Language Teaching Research*, *4*(3), 275–300. https://doi.org/10.1177/136216880000400305

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, UK: Cambridge University Press.

Du, Y., & Wright, B. D. (1997). Effects of student characteristics in a large-scale direct writing assessment. In M. Wilson, G. Engelhard, Jr., & K. Draney (Eds.), *Objective Measurement: Theory into Practice* (Vol. 4, pp. 1–24). Greenwich, CT: Ablex.

Du, Y., Wright, B. D., & Brown, W. L. (1996). *Differential facet functioning detection in direct writing assessment.* Paper presented at the Annual Conference of the American Educational Research Association, New York, NY. Retrieved from http://eric.ed.gov/?id=ED400293

Dufva, M., & Voeten, M. J. (1999). Native language literacy and phonological memory as prerequisites for learning English as a foreign language. *Applied Psycholinguistics*, *20*(03), 329–348. https://doi.org/10.1017/s014271649900301x

Duquette, L., Renié, D., & Laurier, M. (1998). The evaluation of vocabulary acquisition when learning French as a second language in a multimedia environment. *Computer Assisted Language Learning*, *11*(1), 3–34. https://doi.org/10.1076/call.11.1.3.5725

Eccles, J. S. (1983). Expectancies values and academic behaviors. In J. T. Spence (Ed.), *Achievement and Achievement Motives: Psychological and Sociological Approaches* (pp. 78–146). San Francisco, CA: W.H. Freeman and Company.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Educational Testing Service. (2013). *TOEFL Primary Speaking test*. Retrieved from http://www.ets.org/toefl_primary

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing*, *19*(4), 347–368. https://doi.org/10.1191/0265532202lt235oa

Ellis, R. (2008). *Principles of instructed second language acquisition*. Washington, DC: Center for Applied Linguistics.

ELPA21. (n.d.). Retrieved from http://www.elpa21.org/

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037/0033-295x.87.3.215

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. Cambridge, MA: MIT Press.

Espinosa, L. (2012). Assessment of young English-language learners. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0057

Eyben, F. (2016). *Real-time speech and music classification by large audio feature space extraction*. https://doi.org/10.1007/978-3-319-27299-3

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., … Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/taffc.2015.2457417

Eyben, F., Weninger, F., Gross, F., & Schuller, B. W. (2013, October). *Recent developments in openSMILE, the Munich open-source multimedia feature extractor*. Presented at the 21st ACM International Conference on Multimedia, Barcelona, Spain. https://doi.org/10.1145/2502081.2502224

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Field, J. (2018). The cognitive validity of tests of listening and speaking designed for young learners. In N. Saville & C. J. Weir (Eds.), *Examining Young Learners: Research and Practice in Assessing the English of School-Age Learners* (pp. 128–200). New York, NY: Cambridge English.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models* (pp. 131–155). https://doi.org/10.1007/978-1-4612-4230-7_8

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction*, *43*, 1–4. https://doi.org/10.1016/j.learninstruc.2016.02.002

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 763–782). https://doi.org/10.1007/978-1-4614-2018-7_37

French, L. M. (2003). *Phonological working memory and L2 acquisition: A developmental study of Quebec francophone children learning English* (Doctoral dissertation). Universite Laval, Quebec, Canada.

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics*, *29*(3), 463–487. https://doi.org/10.1017/s0142716408080211

Fryer, L. K., Ainley, M., & Thompson, A. (2016). Modelling the links between students' interest in a domain, the tasks they experience and their interest in a course: Isn't interest what university is all about? *Learning and Individual Differences*, *50*, 157–165. https://doi.org/10.1016/j.lindif.2016.08.011

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, *95*(1), 148–162. https://doi.org/10.1037/0022-0663.95.1.148

Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89–119. https://doi.org/10.1080/15434300801934702

Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. Baltimore, MD: Edward Arnold.

Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second-language learning*. Rowley, MA: Newbury House.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*(2), 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Getman, E. P., Cho, Y., & Luce, C. (2016). *Effects of printed option sets on listening item performance among young English-as-a-foreign-language learners* (Research Memorandum No. RM-16-16). Princeton, NJ: Educational Testing Service.

Gilakjani, A. P. (2012). The significant role of multimedia in motivating EFL learners' interest in English language learning. *International Journal of Modern Education and Computer Science*, *4*(4), 57–66. https://doi.org/10.5815/ijmecs.2012.04.08

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*(4), 351–373. https://doi.org/10.1111/j.1745-3984.2005.00020.x

Graddol, D. (2006). *English next*. Retrieved from https://englishagenda.britishcouncil.org

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist*, *50*(1), 14–30. https://doi.org/10.1080/00461520.2014.989230

Grotjahn, R. (1987). On the methodological basis of introspective methods. In C. Faerch & G. Kasper (Eds.), *Introspection in Second Language Research* (pp. 54–81). Clevedon, UK: Multilingual Matters.

Gupta, R., Bone, D., Lee, S., & Narayanan, S. (2016). Analysis of engagement behavior in children during dyadic interactions using prosodic cues. *Computer Speech & Language*, *37*, 47–66. https://doi.org/10.1016/j.csl.2015.09.003

Halliday, M. A. K., & Hasan, R. (1989). *Language, context, and text: Aspects of language in a social-semiotic perspective* (2nd ed.). Hong Kong: Oxford University Press.

Hasselgreen, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing*, *17*(2), 261–277. https://doi.org/10.1191/026553200669937287

Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing*, *22*(3), 337–354. https://doi.org/10.1191/0265532205lt312oa

Hasselgreen, A., & Caudwell, G. (2016). *Assessing the language of young learners*. Sheffield, UK: Equinox Publishing.

Hauck, M. C., Pooler, E., Wolf, M. K., Lopez, A. A., & Anderson, D. P. (2017). Designing task types for English language proficiency assessments for K–12 English learners in the U.S. In M. K. Wolf & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners* (pp. 79–95). https://doi.org/10.4324/9781315674391-5

Hauck, M. C., Wolf, M. K., & Mislevy, R. (2013). *Creating a next-generation system of K–12 English learner (EL) language proficiency assessments*. Princeton, NJ: Educational Testing Service.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. W. He (Eds.), *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency* (pp. 1–25). Amsterdam, The Netherlands: John Benjamins.

Heining-Boynton, A. L., & Haitema, T. (2007). A ten-year chronicle of student attitudes toward foreign language in the elementary school. *The Modern Language Journal*, *91*(2), 149–168. https://doi.org/10.1111/j.1540-4781.2007.00538.x

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*, 36–53. https://doi.org/10.1016/j.compedu.2015.09.005

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

Hsieh, Y.-H., Lin, Y.-C., & Hou, H.-T. (2014). Exploring elementary-school students' engagement patterns in a game-based learning environment. *Educational Technology & Society*, *18*(2), 336–348.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, *102*(4), 880–895. https://doi.org/10.1037/a0019506

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 538–588). Malden, MA: Blackwell Publishing.

IBM Corp. (2015). IBM SPSS Statistics for Windows (Version 23.0). Armonk, NY.

Inbar-Lourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), *The Age Factor and Early Language Learning* (pp. 83–96). https://doi.org/10.1515/9783110218282.83

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*, 401–436. https://doi.org/10.1111/0023-8333.00160

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. https://doi.org/10.1037//0022-3514.79.6.995

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing*, *16*(4), 426–456. https://doi.org/10.1191/026553299667813224

Johnson, Jacqueline S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99. https://doi.org/10.1016/0010-0285(89)90003-0

Johnson, Jeff S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485–505. https://doi.org/10.1177/0265532209340186

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners.* (Technical Report No. 44). Retrieved from University of Minnesota, National Center on Educational Outcomes website: http://eric.ed.gov/?id=ED495909

Johnstone, R. (2000). Context-sensitive assessment of modern languages in primary (elementary) and early secondary education: Scotland and the European experience. *Language Testing*, *17*(2), 123–143. https://doi.org/10.1177/026553220001700202

Jones, B. D. (2009). Motivating students to engage in learning: The MUSIC model of academic motivation. *International Journal of Teaching and Learning in Higher Education*, *21*(2), 272–285.

Jones, L. C., & Plass, J. L. (2002). Supporting listening comprehension and vocabulary acquisition in French with multimedia annotations. *The Modern Language Journal*, *86*(4), 546–561. https://doi.org/10.1111/1540-4781.00160

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631. https://doi.org/10.2307/2285946

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, *1*(4), 381–412. https://doi.org/10.1037//1528-3542.1.4.381

Kalathottukaren, R. T., Purdy, S. C., & Ballard, E. (2015). Behavioral measures to evaluate prosodic skills: A review of assessment tools for children and adults. *Contemporary Issues in Communication Science and Disorders*, *42*, 138–154. https://doi.org/10.1044/cicsd_42_s_138

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (Vol. 4, pp. 17–64). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, *10*(3), 2–10. https://doi.org/10.1007/bf02905780

Keller, J. M. (2009). *Motivational design for learning and performance: The ARCS model approach*. New York, NY: Springer.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing*, *34*(1), 23–48. https://doi.org/10.1177/0265532215595666

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment* (Doctoral dissertation). Teachers College, Columbia University, New York, NY.

Kim, J., & Truong, K. P. (2016). Automatic analysis of children's engagement using interactional network features. *Workshop on Child Computer Interaction*, 23–28. https://doi.org/10.21437/wocci.2016-4

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3–31. https://doi.org/10.1191/0265532202lt218oa

Kondo-Brown, K. (2004). Investigating interviewer-candidate interactions during oral interviews for child L2 learners. *Foreign Language Annals*, *37*(4), 601–613. https://doi.org/10.1111/j.1944-9720.2004.tb02426.x

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, N.J: Lawrence Erlbaum.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*(1), 1–27. https://doi.org/10.1191/0265532204lt272oa

Krapp, A., Hidi, S., & Renninger, A. (2014). Interest, learning, and development. In A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 3–25). New York, NY: Psychology Press.

Kubanek-German, A. (1998). Primary foreign language teaching in Europe: Trends and issues. *Language Teaching*, *31*(4), 193–205. https://doi.org/10.1017/s0261444800013355

Kunnan, A. J. (2008). Large scale language assessments. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education* (2nd ed., Vol. 7, pp. 135–155). https://doi.org/10.1007/978-0-387-30424-3_173

Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* London, UK: Longman.

Lai, E. R. (2011). *Metacognition: A literature review* [Research Report]. Retrieved from Pearson website: https://www.pearsonassessments.com/research

Lambert, C., Philp, J., & Nakamura, S. (2017). Learner-generated content and engagement in second language task performance. *Language Teaching Research*, *21*(6), 665–680. https://doi.org/10.1177/1362168816683559

Lara-Brady, L. G., & Wendler, C. (2013). *A conceptual framework for understanding the use and acquisition of language by English language learners* (Research Memorandum No. RM-13-07). Princeton, NJ: Educational Testing Service.

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research*, *24*(1), 35–63. https://doi.org/10.1177/0267658307082981

Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, *111*(5), 686–705. https://doi.org/10.1037/pspi0000066

Lawson, M. A., & Lawson, H. A. (2013). New conceptual frameworks for student engagement research, policy, and practice. *Review of Educational Research*, *83*(3), 432–479. https://doi.org/10.3102/0034654313480891

Lee, S., & Winke, P. (2017). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing*, *35*(2), 239–269. https://doi.org/10.1177/0265532217704009

Lee, Y.-W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks* (TOEFL Monograph No. MS-28). Princeton, NJ: Educational Testing Service.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23*(2), 131–166. https://doi.org/10.1191/0265532206lt325oa

Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing*, *21*(3), 335–359. https://doi.org/10.1191/0265532204lt287oa

Linacre, J. M. (1999). Facets Rasch measurement computer program (Version 3.22). Chicago, IL: Winsteps.com.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Linacre, J. M. (2017). *A user's guide to Facets: Rasch-model computer program* (3.80.00). Beaverton, OR: Winsteps.com.

Lopriore, L., & Mihaljević Djigunović, J. (2011). Aural comprehension and oral production of young EFL learners. In J. Horváth (Ed.), *Empirical Studies in English Applied Linguistics* (pp. 73–82). Pécs, Hungary: Lingua Franca Csoport.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415–437. https://doi.org/10.1191/0265532205lt303oa

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180. https://doi.org/10.1177/026553229801500202

Macey, W. H., & Schneider, B. (2008). The meaning of employee engagement. *Industrial and Organizational Psychology*, *1*(1), 3–30. https://doi.org/10.1111/j.1754-9434.2007.0002.x

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, *5*(4), 333–369. https://doi.org/10.1016/s0364-0213(81)80017-1

Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2000). Three misconceptions about age and L2 learning. *TESOL Quarterly*, *34*(1), 9–34. https://doi.org/10.2307/3588095

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 31–48). Oxford, UK: Cambridge University.

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction*, *29*, 171–173. https://doi.org/10.1016/j.learninstruc.2013.04.003

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, *90*(2), 312–320. https://doi.org/10.1037/0022-0663.90.2.312

McKay, P. (2006). *Assessing young language learners*. Cambridge, UK: Cambridge University Press.

McNamara, D. S., Jackson, G. T., & Graesser, A. (2010). Intelligent tutoring and games. In Y. K. Baek (Ed.), *Gaming for Classroom-Based Learning: Digital Role Playing as a Motivator of Study* (pp. 44–65). Hershey, PA: IGI Global.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

Meltzer, J., & Hamann, E. (2004). *Meeting the needs of adolescent English language learners for literacy development and content area learning, Part 1: Focus on motivation and engagement*. Providence, RI: The Education Alliance at Brown University.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Mihaljević Djigunović, J. (2009). Individual differences in early language programmes. In M. Nikolov (Ed.), *The Age Factor and Early Language Learning* (pp. 199–226). Berlin, Germany: Mouton de Gruyter.

Mihaljević Djigunović, J., & Krevelj, S. L. (2009). Instructed early SLA: Development of attitudes. *Studia Romanica et Anglica Zagrabiensia*, *54*, 137–156.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496. https://doi.org/10.1191/0265532202lt241oa

Mohsen, M. A., & Balakumar, M. (2011). A review of multimedia glosses and their effects on L2 vocabulary acquisition in CALL literature. *ReCALL*, *23*(2), 135–159. https://doi.org/10.1017/s095834401100005x

Moyer, A. (1999). Ultimate attainment in L2 phonology. *Studies in Second Language Acquisition*, *21*(1), 81–108. https://doi.org/10.1017/s0272263199001035

Mozgalina, A. (2015). More or less choice? The influence of choice on task motivation and task engagement. *System*, *49*, 120–132. https://doi.org/10.1016/j.system.2015.01.004

Muñoz, C. (2009). Input and long-term effects of early learning in a formal setting. In M. Nikolov (Ed.), *The Age Factor and Early Language Learning* (pp. 141–159). Berlin, Germany: Mouton de Gruyter.

Muñoz, C. (2012). Age-appropriate instruction and assessment for school-age learners. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0015

Muthén, L. K., & Muthén, B. O. (2017). Mplus (Version 8). Los Angeles, CA: Muthén & Muthén.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (Research Report No. 65). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

National Center for Education Statistics. (2004). *English language learner students in U.S. public schools: 1994 and 2000*. Retrieved from http://nces.ed.gov/pubs2004/2004035.pdf

Nikolov, M. (1999). 'Why do you learn English?' 'Because the teacher is short.' A study of Hungarian children's foreign language learning motivation. *Language Teaching Research*, *3*(1), 33–56. https://doi.org/10.1177/136216889900300103

Oga-Baldwin, W. L. Q., Nakata, Y., Parker, P., & Ryan, R. M. (2017). Motivating young language learners: A longitudinal model of self-determined motivation in elementary school foreign language classes. *Contemporary Educational Psychology*, *49*, 140–150. https://doi.org/10.1016/j.cedpsych.2017.01.010

Oh, S. R. (2018). *Investigating test-takers' use of linguistic tools in second language academic writing assessment* (Doctoral dissertation). Teachers College, Columbia University, New York, NY.

Oliver, R. (1998). Negotiation of meaning in child interactions. *The Modern Language Journal*, *82*(3), 372–386. https://doi.org/10.2307/329962

Oller, J. W. (1979). *Language tests in schools: A pragmatic approach*. London, UK: Longman.

Oller, J. W. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller (Ed.), *Issues in Language Testing Research* (pp. 3–10). Rowley, MA: Newbury House.

Owens, M., Stevenson, J., Norgate, R., & Hadwin, J. A. (2008). Processing efficiency theory in children: Working memory as a mediator between trait anxiety and academic performance. *Anxiety, Stress & Coping*, *21*(4), 417–430. https://doi.org/10.1080/10615800701847823

Papp, S. (2018). Criterion-related validity of tests of English for young learners. In N. Saville & C. J. Weir (Eds.), *Examining Young Learners: Research and Practice in Assessing the English of School-Age Learners*. New York, NY: Cambridge English.

Park, S. (2008). *An exploration of examinee abilities, rater performance, and task differences using diverse analytic techniques* (Doctoral dissertation). University of Hawaii, Honolulu, HI.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2010). Innovative items for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 215–230). https://doi.org/10.1007/978-0-387-85461-8_11

Payán, R. M., & Nettles, M. T. (2008). *Current state of English-language learners in the US K–12 student population*. Princeton, NJ: Educational Testing Service.

Pearson. (2009). *PTE Young Learners handbook*. Retrieved from http://www.pearsonpte.com/PTEYoungLearners

Pérez, J. M., Gálvez, R. H., & Gravano, A. (2016). Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement. *INTERSPEECH 2016*, 1270–1274. https://doi.org/10.21437/interspeech.2016-587

Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics*, *36*, 50–72. https://doi.org/10.1017/s0267190515000094

Phung, L. (2017). Task preference, affective response, and engagement in L2 use in a US university context. *Language Teaching Research*, *21*(6), 751–766. https://doi.org/10.1177/1362168816683561

Piaget, J. (1964). Development and learning. In R. E. Ripple & V. N. Rockcastle (Eds.), *Piaget Rediscovered: Selected Papers From a Report of the Conference of Cognitive Studies and Curriculum Development* (pp. 7–20). Ithaca, NY: Cornell University.

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners*. Princeton, NJ: Educational Testing Service.

Poehner, M. E., Zhang, J., & Lu, X. (2017). Computerized dynamic assessments for young language learners. In M. K. Wolf & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners* (pp. 214–233). https://doi.org/10.4324/9781315674391-12

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing*, *4*(1), 72–92. https://doi.org/10.1177/026553228700400107

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Purpura, J. E. (2017). Assessing meaning. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and Assessment* (pp. 33–61). https://doi.org/10.1007/978-3-319-02261-1_1

QSR International. (2018). NVivo qualitative data analysis software (Version 12). Retrieved from https://www.qsrinternational.com/nvivo/home

Rahman, T., & Mislevy, R. J. (2017). *Integrating cognitive views into psychometric models for reading comprehension assessment: Integrating cognitive views into psychometric models* (Research Report No. RR-17-35). https://doi.org/10.1002/ets2.12163

Rea-Dickins, P. (2000). Current research and professional practice: Reports of work in progress into the assessment of young language learners. *Language Testing*, *17*(2), 245–249. https://doi.org/10.1177/026553220001700207

Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 3–19). https://doi.org/10.1007/978-1-4614-2018-7_1

Rigby, S., & Ryan, R. M. (2011). *Glued to games: How video games draw us in and hold us spellbound*. Santa Barbara, CA: Praeger.

Rixon, S. (2013). *British council survey of policy and practice in primary English language.* London, UK: British Council.

Rixon, S. (2018). Consequential validity of tests of English for young learners: The impact of assessment on young learners. In N. Saville & C. J. Weir (Eds.), *Examining Young Learners: Research and Practice in Assessing the English of School-Age Learners* (pp. 547–587). New York, NY: Cambridge English.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, *43*(1), 1–32. https://doi.org/10.1515/iral.2005.43.1.1

Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching*, *45*(3). https://doi.org/10.1515/iral.2007.007

Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, *17*(1), 20–32.

Ryan, C. (2013). *Language use in the United States: 2011* (No. ACS-22). Retrieved from U.S. Census Bureau website: https://www.census.gov

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. New York, NY: Guilford Press.

Sang, F., Schmitz, B., Volmer, H., Baumert, J., & Roeder, P. (1986). Models of second language competence: A structural equation modeling approach. *Language Testing*, *3*(1), 54–79. https://doi.org/10.1177/026553228600300103

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243–248. https://doi.org/10.1007/s11336-009-9135-y

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample* (Research Report No. TOEFLiBT-04). https://doi.org/10.1002/j.2333-8504.2008.tb02095.x

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 005–030. https://doi.org/10.1177/0265532208097335

Scalise, K. (2012, May). *Using technology to assess hard-to-measure constructs in the common core state standards and to expand accessibility*. Presented at the Invitational Research Symposium on Technology Enhanced Assessments, Washington, DC.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465–493. https://doi.org/10.1177/0265532208094273

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227–256. https://doi.org/10.1016/s0167-6393(02)00084-5

Scherer, K. R. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language*, *27*(1), 40–58. https://doi.org/10.1016/j.csl.2011.11.003

Schmidgall, J. E. (2017). *The consistency of TOEIC Speaking scores across ratings and tasks* (Research Report No. RR-17-46). https://doi.org/10.1002/ets2.12178

Schuller, B. W., Steidl, S., & Batliner, A. (2009, September). *The INTERSPEECH 2009 emotion challenge*. Presented at the Conference of the International Speech Communication Association, Brighton, UK.

Schuller, B. W., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. S. (2010, September). *The INTERSPEECH 2010 paralinguistic challenge*. Presented at the Conference of the International Speech Communication Association, Makuhari, Japan.

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, *50*(5), 489–499. https://doi.org/10.1002/pits.21689

Service, E. (1992). Phonology, working memory, and foreign-language learning. *The Quarterly Journal of Experimental Psychology*, *45*(1), 21–50. https://doi.org/10.1080/14640749208401314

Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, *16*(02), 155–172. https://doi.org/10.1017/s0142716400007062

Shaaban, K. (2001). Assessment of young learners. *English Teaching Forum*, *39*(4), 16–23.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405–450.

Shepard, L. A. (1994). The challenges of assessing young children appropriately. *The Phi Delta Kappan*, *76*(3), 206–212.

Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology*, *71*(3), 549. https://doi.org/10.1037/0022-3514.71.3.549

Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (2009). Engagement and disaffection as organizational constructs in the dynamics of motivational development. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of Motivation at School* (pp. 223–245). New York, NY: Routledge.

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*(4), 435–464. https://doi.org/10.1177/0265532208094272

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, *50*(3), 345–362.

Spolsky, B. (1973). What does it mean to know a language; or how do you get someone to perform his competence? In J. W. Oller & J. Richards (Eds.), *Focus on the Learner: Pragmatic Perspectives of the Language Teacher* (pp. 64–176). Rowley, MA: Newbury House.

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL Internet-based test across subgroups* (Research Report No. TOEFLiBT-07). https://doi.org/10.1002/j.2333-8504.2008.tb02152.x

Svalberg, A. M.-L. (2009). Engagement with language: Interrogating a construct. *Language Awareness*, *18*(3–4), 242–258. https://doi.org/10.1080/09658410903197264

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Taylor, L., & Saville, N. (2002). Developing English language tests for young learners. *Cambridge ESOL: Research Notes*, *7*, 2–5.

Teasdale, A., & Leung, C. (2000). Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing*, *17*(2), 163–184. https://doi.org/10.1191/026553200672041197

Teixeira, J. P., & Fernandes, P. O. (2014). Jitter, shimmer and HNR classification within gender, tones and vowels in healthy voices. *Procedia Technology*, *16*, 1228–1237. https://doi.org/10.1016/j.protcy.2014.10.138

Thurman, J. (2013). Choice and its influence on intrinsic motivation and output in task-based language teaching. *The Asian EFL Journal Quarterly*, *15*(1), 202–245.

Traphagan, T. W. (1997). Interviews with Japanese FLES students: Descriptive analysis. *Foreign Language Annals*, *30*(1), 98–110. https://doi.org/10.1111/j.1944-9720.1997.tb01320.x

Truesdale, D. M., & Pell, M. D. (2018). The sound of passion and indifference. *Speech Communication*, *99*, 124–134. https://doi.org/10.1016/j.specom.2018.03.007

Trushell, J., Burrell, C., & Maitland, A. (2001). Year 5 pupils reading an “interactive storybook” on CD-ROM: Losing the plot? *British Journal of Educational Technology*, *32*(4), 389–401. https://doi.org/10.1111/1467-8535.00209

Turkan, S., & Adler, R. (2011). *Conceptual framework for the assessment of young learners of English as a foreign language* [Unpublished manuscript]. Princeton, NJ: Educational Testing Service.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment* (pp. 255–273). Boston, MA: De Gruyter Mouton.

Vanpatten, B., & Cadierno, T. (1993). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal*, *77*(1), 45–57. https://doi.org/10.1111/j.1540-4781.1993.tb01944.x

Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist*, *41*(1), 19–31. https://doi.org/10.1207/s15326985ep4101_4

Verhallen, M. J. A. J., & Bus, A. G. (2009). Video storybook reading as a remedy for vocabulary deficits: Outcomes and processes. *Journal for Educational Research Online*, *1*(1), 172–196.

Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167–197. https://doi.org/10.1191/0265532206lt326oa

WIDA. (n.d.). Retrieved from https://www.wida.us/

Wiersma, W., & Jurs, S. G. (2008). *Research methods in education: An introduction* (9th ed.). Boston, MA: Pearson.

Winke, P. (2015, October). *Investigations into language assessment using eye-tracking methods*. Plenary address presented at the East Coast Organization of Language Testers Conference, Washington, DC.

Winke, P., Gass, S., & Myford, C. (2011). *The relationship between raters' prior language study and the evaluation of foreign language speech samples* (Research Report No. TOEFLiBT-16). https://doi.org/10.1002/j.2333-8504.2011.tb02266.x

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Wolf, M. K., & Butler, Y. G. (2017). An overview of English language proficiency assessments for young learners. In M. K. Wolf & Y. G. Butler (Eds.), *English Language Proficiency Assessments for Young Learners* (pp. 3–21). https://doi.org/10.4324/9781315674391-1

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement: What Every Psychologist and Educator Should Know* (pp. 65–104). Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: Mesa Press.

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, *13*(2), 48–67.

Yu, C., Aoki, P., & Woodruff, A. (2004). *Detecting user engagement in everyday conversations*. Presented at the International Conference on Spoken Language Processing, Jeju Island, South Korea.

Zangl, R. (2000). Monitoring language skills in Austrian primary (elementary) schools: A case study. *Language Testing*, *17*(2), 250–260. https://doi.org/10.1177/026553220001700208

Appendix A

Pre-Test Demographic Questionnaire

Item	Options
Country:	Egypt Bahrain Brazil China Gaza Iraq Japan Jordan Jordan Korea Kuwait Lebanon Macedonia Morocco Peru Qatar Saudi Arabia Syria Tunisia UAE Vietnam West Bank Yemen
Date of birth:	(month, day, year)
Native language:	Arabic Berber Chinese English Japanese Korean Kurdish Persian Portuguese Spanish Turkish Vietnamese Other

Item	Options
Gender:	Female Male
At my school I am in:	Grade 1 Grade 2 Grade 3 Grade 4 Grade 5 Grade 6 Other
I have studied English for:	Less than 1 year 1 year 2 years 3 years 4 years 5 years 6 years or more
Each week, at my regular school, I have _____ of English classes.	0 hours 1 hour 2 hours 3 hours 4 hours 5 hours or more
Each week, at my after-school program, I have _____ of English classes.	0 hours 1 hour 2 hours 3 hours 4 hours 5 hours or more
Each week, outside the classroom, I study English for _____.	0 hours 1 hour 2 hours 3 hours 4 hours 5 hours or more

Appendix B

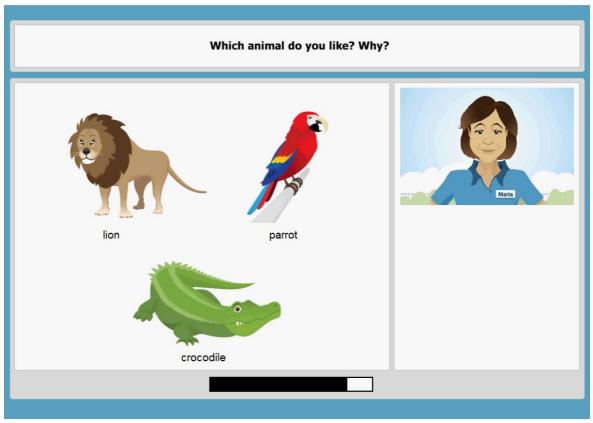
TOEFL Primary Speaking Tasks

Following are descriptions of the 11 *TOEFL Primary* Speaking tasks used in the current study.

Task 1: What's your favorite animal?

Task type: Express an Opinion

Description: The test taker is presented with the picture, text, and spoken audio of three animals: a lion, a parrot, and a crocodile. Then Maria, the zookeeper, asks the test taker what his or her favorite animal is, and why. The test taker must give and explain his or her opinion.



Task 2: How do you feed the birds?

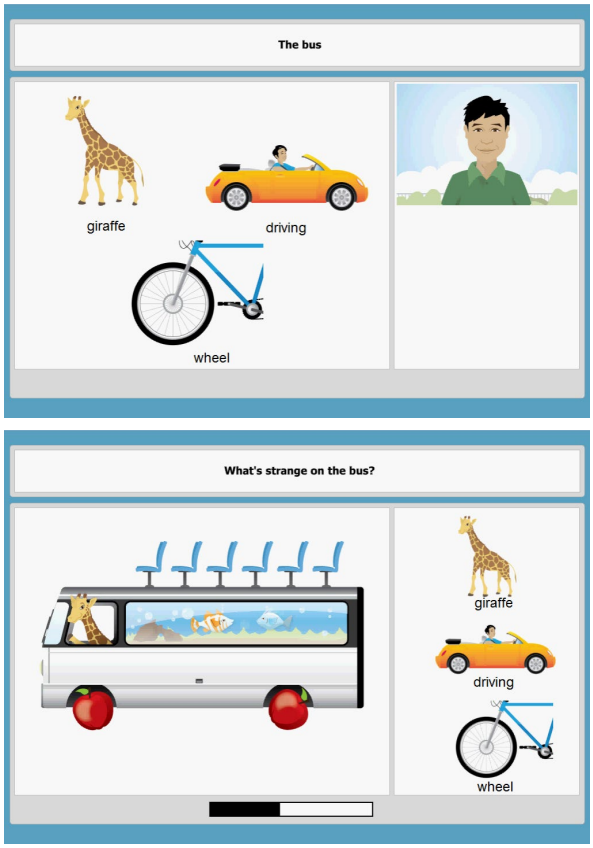
Task type: Give Directions

Description: The test taker is presented with a four-picture sequence of a boy feeding the birds. The four pictures are numbered, and they are presented one at a time so that initially only the first picture is seen, then the first two pictures, then the first three, and so on. Then Becca, a peer, reports that she too wants to feed the birds, but she didn't see what to do. The test taker must explain the directions to Becca.

Task 3: What's strange on the bus?

Task type: Describe a Picture

Description: The test taker is presented with three words that will be useful in the response: giraffe, driving, and wheel. Then an unusual bus is presented, unusual in that it features seats on the roof, a giraffe driving, apples for wheels, and an aquarium through the windows. The test taker must describe what is strange about the bus.



Task 4

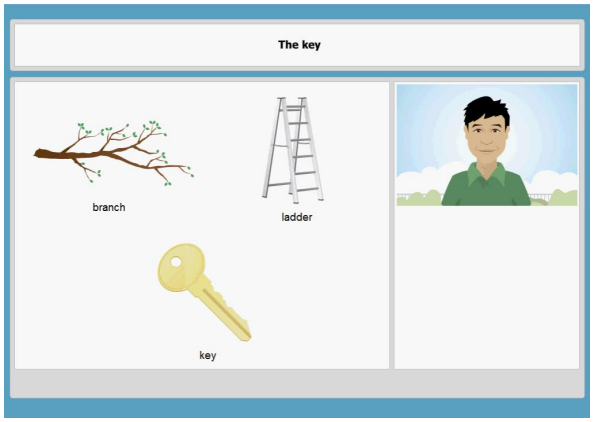
Task type: Describe a Picture

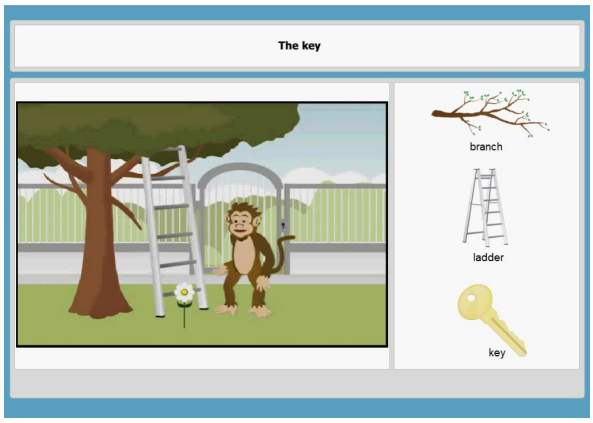
Description: This task tests the same speech function as the previous task, but the task content has not been released to the public by Educational Testing Service (ETS).

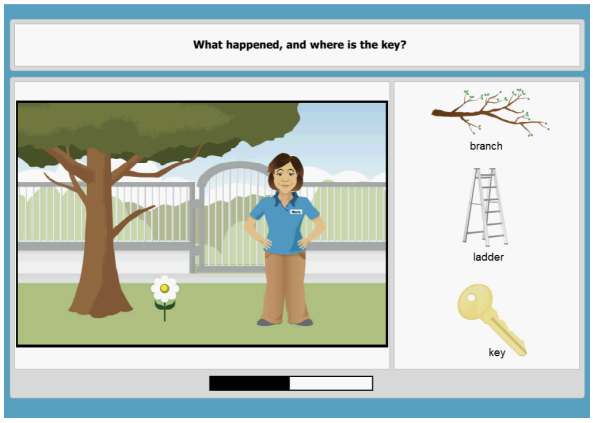
Task 5: What happened to the key?

Task type: Retell a Story

Description: The test taker is presented with three words that will be useful in the response: branch, ladder, and key. Then a video animation is presented showing Billy the monkey hiding the key to the gate. Billy pulls on the branch to reveal a ladder. He climbs the ladder and puts the key in the tree. When he gets back on the ground, he pulls on a flower to retract the ladder. The video is presented twice to minimize memory effects. Then Maria, the zookeeper, explains that she needs the key and asks the test taker to explain what just happened. The test taker must retell the events from the animation.



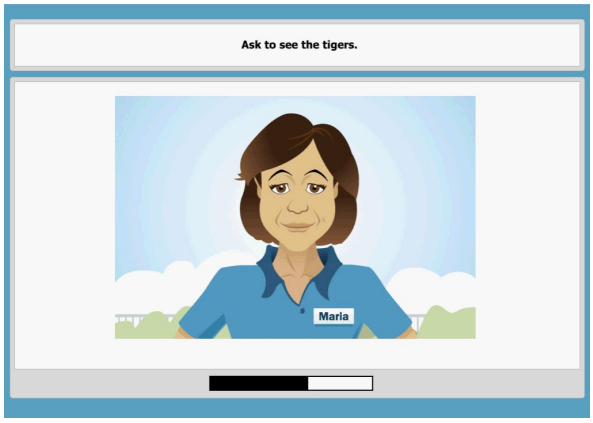




Task 6: Ask to see the tigers.

Task type: Ask Questions

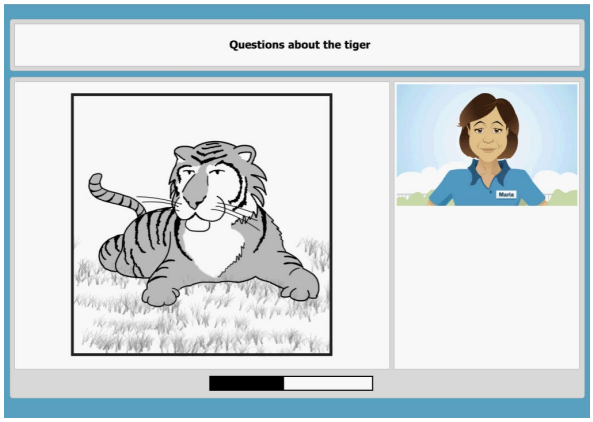
Description: Becca, a virtual peer, tells the test taker to ask Maria, the zookeeper, if they can go to see the tigers. The test taker must make the request of Maria.



Task 7: Ask three questions about the tiger.

Task type: Ask Questions

Description: The test taker is invited by Maria, the zookeeper, to ask questions about the tiger. Maria first gives some examples of things that could be asked about. Then the test taker must ask three questions about the tiger.



Task 8

Task type: Retell a Story

Description: This task tests the same speech function as Task 5, but the task content has not been released to the public by ETS.

Task 9

Task type: Ask Questions

Description: This task tests the same speech function as Task 6, but the task content has not been released to the public by ETS.

Task 10

Task type: Give Directions

Description: This task tests the same speech function as Task 2, but the task content has not been released to the public by ETS.

Task 11

Task type: Express an Opinion

Description: This task tests the same speech function as Task 1, but the task content has not been released to the public by ETS.

Appendix C

Rubrics

0-to-3-Point Scoring Guide

For the following communication goals: **express basic emotions, feelings, and opinions**
give simple descriptions
make simple requests
ask questions

Score	Language Use, Content, and Delivery Descriptors
3	**The test taker achieves the communication goal.** A typical response at the 3 level is characterized by the following. • The meaning is clear. Minor errors in grammar or word choice do not affect task achievement. • The response is accurate and complete, and the content is appropriate for the task. • Speech is intelligible, and the delivery is generally fluid. It requires minimal listener effort for comprehension.
2	**The test taker partially achieves the communication goal.** A typical response at the 2 level is characterized by the following. • The meaning is mostly clear. Some errors in grammar or word choice may interfere with task achievement. • The response is not fully accurate or complete, or the content is not fully appropriate for the task. • Speech is generally intelligible, but the delivery may be slow, choppy, or hesitant. It requires some listener effort for comprehension.
1	**The test taker attempts to achieve the communication goal.** A typical response at the 1 level is characterized by the following. • The meaning is obscured because of frequent errors in grammar and word choice. • The response is inaccurate or incomplete, or the content is inappropriate for the task. • Speech is mostly unintelligible or unsustained. It requires significant listener effort for comprehension.
0	The test taker does not attempt to achieve the communication goal OR the response contains no English OR the response is off topic and does not address the prompt.

0-to-5-Point Scoring Guide

For the following communication goals: **explain and sequence simple events**
give directions

Score	Language Use, Content, and Delivery Descriptors
5	**The test taker fully achieves the communication goal.** A typical response at the 5 level is characterized by the following. • The meaning is clear. Grammar and word choice are effectively used. Minor errors do not affect task achievement. Coherence may be assisted by use of connecting devices. • The response is full and complete. Events are described accurately and are easy to follow. • Speech is fluid with a fairly smooth, confident rate of delivery. It contains few errors in pronunciation and intonation. It requires little or no listener effort for comprehension.
4	**The test taker achieves the communication goal.** A typical response at the 4 level is characterized by the following. • The meaning is mostly clear. Some errors in grammar and word choice may be noticeable, but the task is still achieved. Use of connecting devices to link ideas may be limited. • The response is mostly complete. Descriptions contain minor lapses or inaccuracies, but the events can still be readily followed. • Speech is mostly fluid and sustained, though some hesitation and choppiness may occur. It contains minor errors in pronunciation and intonation. It requires minimal listener effort for comprehension.
3	**The test taker partially achieves the communication goal.** A typical response at the 3 level is characterized by the following. • The meaning is sometimes obscured. Errors in grammar and word choice are noticeable and limit task achievement. The response may include attempts to use connecting devices. • The response is somewhat complete. Lapses and inaccuracies require the listener to fill in the gaps between key events. • Speech may be sustained throughout, but the pace may be slow, choppy, or hesitant. It contains errors in pronunciation and intonation. It requires some listener effort for comprehension.

Score	Language Use, Content, and Delivery Descriptors
2	**The test taker is limited in achieving the communication goal.** A typical response at the 2 level is characterized by the following. • The meaning is often obscured because of errors in grammar and word choice. Attempts at using connecting devices are unsuccessful or absent. • The response is mostly incomplete. Multiple lapses and gaps make it difficult for listeners unfamiliar with the events to follow along. Meaningful content may be conveyed through repetition. • Speech is noticeably slow, choppy, or hesitant throughout and may include long pauses. It contains frequent errors in pronunciation and intonation. It requires listener effort for comprehension.
1	**The test taker attempts to achieve the communication goal.** A typical response at the 1 level is characterized by the following. • The meaning is obscured because of frequent errors. Grammar and word choice are extremely limited and often inaccurate. • The response is incomplete. Major lapses and gaps make events unclear. The response may consist of a single word or a few words related to the prompt. It may be highly repetitive. • Speech is not sustained or is mostly incomprehensible. It contains numerous errors in pronunciation and intonation. It requires significant listener effort for comprehension.
0	The test taker does not attempt to achieve the communication goal OR the response contains no English OR the response is off topic and does not address the prompt.

Appendix D

Acoustic Features Considered as Possible Indicators of Engagement

This is a list of the 100 acoustic, prosodic, and temporal features that were extracted from audio files of spoken responses and considered as possible indicators of engagement. Each feature is characterized by a low-level descriptor and a functional. Features extracted by Praat (Version 6.04; Boersma & Weenink, 2016) are marked with asterisks (*). The remaining features were extracted as part of the eGemaps feature set (Eyben et al., 2016) using OpenSmile (Version 2.30; Eyben et al., 2013). For more detailed explanations of these features and how they are measured, please consult Eyben (2016).

Low-Level Descriptors	Functionals	Comments
Syllables	Sum*	These were counted using a script to detect syllable nuclei (De Jong & Wempe, 2009) that was updated in 2010 by Quené, Persoon, and De Jong.
Time on task	Sum*	This includes the duration from the start of the response time until the examinee stops speaking.
Speech rate	Mean*	This was calculated by dividing the number of syllables by the time on task.
Speaking time	Sum*	This includes the duration from when the examinee starts speaking until the examinee stops speaking.
Speaking rate	Mean*	This was calculated by dividing the number of syllables by the speaking time.
Phonation time	Sum*	This includes only the time when sounds are actually voiced.
Articulation rate	Mean*	This was calculated by dividing the number of syllables by the phonation time.
Initiation time	Sum*	This includes the duration from the start of the response time until the examinee starts speaking (i.e., time on task – speaking time).
Syllables per run	Mean*	A run is the voiced segment between pauses.
Voiced segments per second	Mean	

Low-Level Descriptors	Functionals	Comments
Length of voiced segments	Mean, standard deviation	
Length of unvoiced segments	Mean, standard deviation	
F_0	Mean, standard deviation, minimum*, 20th percentile, 50th percentile, 80th percentile, maximum*, range*, range from 20th percentile to 80th percentile	F_0 is the fundamental frequency, which is the acoustic equivalent of pitch.
F_0 when the slope was rising	Mean, standard deviation	
F_0 when the slope was falling	Mean, standard deviation	
Jitter	Mean, standard deviation	Jitter refers to the moment-to-moment perturbations in F_0.
Energy	Mean, standard deviation, 20th percentile, 50th percentile, 80th percentile, range from 20th percentile to 80th percentile	Energy is synonymous with intensity or loudness.
Energy when the slope was rising	Mean, standard deviation	
Energy when the slope was falling	Mean, standard deviation	
Equivalent sound level	Mean	This is derived from the root mean square.
Shimmer	Mean, standard deviation	Shimmer refers to the moment-to-moment perturbations in energy.
Energy peaks per second	Mean	
F_1 frequency	Mean, standard deviation	F_1 is the first formant, which is determined by the height of the tongue.
F_1 bandwidth	Mean, standard deviation	
F_1 amplitude	Mean, standard deviation	Measures of amplitude (energy) are relative to that of F_0.
F_2 frequency	Mean, standard deviation	F_2 is the second formant, which is determined by how forward or back the tongue is.
F_2 bandwidth	Mean, standard deviation	
F_2 amplitude	Mean, standard deviation	
F_3 frequency	Mean, standard deviation	F_3 is the third formant, which is related to the shape of the lips.
F_3 bandwidth	Mean, standard deviation	
F_3 amplitude	Mean, standard deviation	
Harmonicity	Mean, standard deviation	Harmonicity is the logarithmic harmonic-to-noise ratio.
Harmonic difference H1-H2	Mean, standard deviation	This is the ratio of the energy of the 1st F_0 harmonic to the energy of the 2nd F_0 harmonic.

Low-Level Descriptors	Functionals	Comments
Harmonic difference H1-A3	Mean, standard deviation	This is the ratio of the energy of the 1st F_0 harmonic to the energy of the highest F_3 harmonic.
Spectral flux	Mean, standard deviation	Flux refers to the difference of the spectra of two consecutive frames.
MFCC 1	Mean, standard deviation	MFCC refers to the Mel-Frequency Cepstral Coefficient.
MFCC 2	Mean, standard deviation	
MFCC 3	Mean, standard deviation	
MFCC 4	Mean, standard deviation	
Alpha ratio (voiced)	Mean, standard deviation	This is the ratio of the summed energies from 50–1,000 Hertz (Hz) and 1,000–5,000 Hz.
Hammarberg index (voiced)	Mean, standard deviation	This is the ratio of the strongest energy peak between 0 and 2,000 Hz to the strongest energy peak between 2,000 and 5,000 Hz.
Spectral slope, 0-500 Hz (voiced)	Mean, standard deviation	This is a measure of how quickly the spectrum of a sound tails off towards high frequencies.
Spectral slope, 500-1500 Hz (voiced)	Mean, standard deviation	
Spectral flux (voiced)	Mean, standard deviation	
MFCC 1 (voiced)	Mean, standard deviation	
MFCC 2 (voiced)	Mean, standard deviation	
MFCC 3 (voiced)	Mean, standard deviation	
MFCC 4 (voiced)	Mean, standard deviation	
Alpha ratio (unvoiced)	Mean	
Hammarberg index (unvoiced)	Mean	
Spectral slope, 0-500 Hz (unvoiced)	Mean	
Spectral slope, 500-1500 Hz (unvoiced)	Mean	
Spectral flux (unvoiced)	Mean	

Appendix E

Script for Child Interviews

We want to learn how to make English tests more fun and interesting for young students like you. First, you will take a short speaking test on the computer. How well you speak in English is not important, and we will not give you a grade. After the test, we will ask you some questions about it. Please tell us as much as possible about what you think. There are no wrong answers. What you say is really important. We are going to use a video camera so we don't forget anything. We will not use your name or share your image with anyone. What are we going to do today?

[Understandings are confirmed and clarified as needed.]

Okay. Sign here.

We are going to start by watching the introduction to the test.

[The test introduction is played.]

So, when you talk, remember to speak loudly and clearly. Now you are going to do four speaking activities. When you are finished, I'm going to ask you some questions about them.

[Tasks 1, 3, 5, and 7 are administered.]

So, what did you think of the speaking activities?

There were four activities: The first was "What's your favorite animal." The second was "What's strange on the bus?" The third was "What happened to the key?" And the fourth was "Ask three questions about the tiger." *[Screenshots of the activities are shown to remind the participant what was just seen.]* Which activity was your favorite? Why?

Which question did you like the least? Why?

What would make it more fun and interesting?

This activity gave you a choice of things you could talk about. *[The screenshot for Task 1 is shown.]* You could choose to talk about a lion, a parrot, or a crocodile. What did you think about having a choice like that? Did this make the activity fun and interesting? Why?

This activity gave you some words at the beginning to help you. *[The screenshot for Task 3 is shown.]* It gave you words for giraffe, driving, and wheel. What did you think about getting some words to help you? Did this make the activity fun and interesting? Why?

This activity gave you a situation that would not happen in real life. *[The screenshot for Task 5 is shown.]* It showed you a bus with a giraffe driving it, with apples for wheels, and with fish in the back. What did you think about having a situation that would not happen in real life? Did this make the activity more fun and interesting? Why?

This activity had a short video showing what Billy the monkey did with the key. *[The screenshot for Task 7 is shown.]* What did you think about having a video like this in the activity? Did this make the activity more fun and interesting? Why?

Can you think of anything else that would help us make these activities more fun and interesting? Tell me more…

What is your birth date?

Thank you for all your help today!

Appendix F

Unstandardized Coefficients for the Final MIMIC Model

Observed indicator variable or unobserved latent variable	Unobserved latent variable or observed causal variable	B	Standard Error
Rater 1 score	Task performance	1.00	
Rater 2 score	Task performance	1.00	0.01
Mean harmonicity	Task engagement	1.00	
Mean shimmer	Task engagement	–0.68	0.03
Standard deviation of shimmer	Task engagement	0.58	0.03
Task performance	Task engagement	0.16	0.01
Task performance	Examinee ability	0.38	0.01
Task performance	Task type difficulty	–0.24	0.03
Task performance	Presence of topical choice	0.10	0.02
Task performance	Presence of vocabulary support	–0.15	0.03
Task performance	Presence of novelty	–0.27	0.04
Task performance	Presence of video animation	–0.10	0.03
Task engagement	Presence of topical choice	0.11	0.03
Task engagement	Presence of novelty	0.35	0.03
Rater 1 score	Rubric type	0.44	0.03
Rater 1 score	Rater 1 severity	–0.32	0.02
Rater 2 score	Rubric type	0.44	0.03
Rater 2 score	Rater 2 severity	–0.32	0.02