TOWARDS OPTIMAL MEASUREMENT AND THEORETICAL GROUNDING OF L2 ENGLISH ELICITED IMITATION: EXAMINING SCALES, (MIS)FITS, AND PROMPT FEATURES FROM ITEM RESPONSE THEORY AND RANDOM FOREST APPROACHES

by

Ji-young Shin

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Department of English
West Lafayette, Indiana
December 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. April Ginther, Chair
College of Liberal Arts

Dr. Anne Traynor
College of Education

Dr. Tony Silva
College of Liberal Arts

Dr. Xun Yan
College of Liberal Arts and Sciences, UIUC

Dr. Lixia Cheng
Purdue Language and Cultural Exchange

Approved by:
Dr. Dorsey Armstrong

Dedicated to My Parents in Heaven

ACKNOWLEDGMENTS

Words cannot express how grateful I am to my advisor and mentor Prof. April Ginther. Throughout my doctoral studies, not to mention for this dissertation, she has provided rigorous guidance with her wealth of knowledge, experience, and insights, as well as a tremendous amount of encouragement, support, and time. Her dedication certainly has gone above and beyond the expectations towards a doctoral advisor. She has been a great example as a great scholar and researcher, insightful language tester, and compassionate mentor, who I will endeavor to emulate in my research and teaching.

I am also sincerely thankful to my committee members, Prof. Anne Traynor, Prof. Tony Silva, Prof. Xun Yan, and Dr. Lixia Cheng. Without the extraordinary insights, guidance, and support they provided, I could not have successfully completed my dissertation.

I am also grateful to Prof. Shelley Staples and Prof. Bradley Dilger. They have provided me tremendous research opportunities, which enriched my dissertation and scholarship in general. Their support and encouragement enabled me to go thorough hardships during my doctoral studies.

My gratitude also goes to Prof. Wayne Wright. His teaching introduced me new perspectives and his support and encouragement for my career meant a lot.

My spiritual father Joshua Cho and my church family have been an amazing source of support. I cannot thank you enough for their prayer, generosity, and love throughout my life, and during my doctoral study, in particular.

I appreciate Zhaozhe Wang for going through this process with me together, with warm support, countless help, and steady love.

I would also like to thank my son, the most precious gift from God, Hyun-gun. He has been my everlasting joy and gratitude ever since he came to my world.

I also give my thanks to my parents in Heaven, Deockyeong Shin and Sangwan Han, who taught me patience, trust, sacrifice, benevolence, and true love. Finally, I give my wholehearted gratitude to my God for his never-ending grace and help, particularly through these wonderful people. I dedicate this dissertation to my beloved parents and my God.

사사

나를 만드신 이, 하나님 아버지께, 이 논문과 박사 과정을 마칠 수 있도록 은혜와 지혜를 부어 주심을 감사합니다. 힘들고 어려운 고비마다 돕는 손길들을 보내주셨습니다.

박사과정 지도교수님 Prof. April Ginther 께 귀한 학문적 가르침을 주시고 인내와 애정으로 아낌없이 멘토링을 해 주심을 감사드립니다. 훌륭한 학자이자 연구자로서 열정적인 멘토로서 큰 귀감이 되어 주셨습니다. 교수님께서 본을 보여주신 가르침들을 향후 연구와 교수, 멘토링에서 실천해 나가리라 다짐합니다.

네 분의 귀한 커미티 교수님들, Prof. Anne Traynor, Prof. Tony Silva, Prof Xun Yan, Dr. Lixia Cheng 의 지도 편달이 없었으면 이 논문은 완성되지 못했을 것입니다.

또한 늘 기도로 사랑으로 성원해 주시는 나의 영적인 아버지 조희서 목사님과 영적인 가족 서울씨티교회 성도님들에게 감사드립니다.

박사과정 내내 도움과 힘이 되어 준 평생의 친구 Zhaozhe Wang 에게 감사합니다.

마지막으로, 사랑하는 아들, 하나님의 가장 귀한 선물, 현건이와, 사랑하고 존경하며 너무나 보고픈 나의 아버지 어머니, 신덕영, 한상완에게 이 논문을 바칩니다.

TABLE OF CONTENTS

PHASE I. MEASURING SEMANTIC AND GRAMMATICAL ACCURACY: OPTIMAL SCALES/ SCORING METHODS AND MISFIT ANALYSIS

PHASE II. BEYOND MEASURING ACCURACY: RELATIONSHIP BETWEEN EI PROMT COMPLEXITY AND EI MEASURMENT OF ACCURACY

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CCC: category characteristic curves

CTT: classical testing theory

EI: elicited imitation

GRM: graded response model

IRT: item response theory

L1: first language

L2: second language

MLM: multi-level modeling

OBB out-of-bag

RF: random forest

ABSTRACT

The present dissertation investigated the impact of scales / scoring methods and prompt linguistic features on the meausrement quality of L2 English elicited imitation (EI). Scales / scoring methods are an important feature for the validity and reliabilty of L2 EI test, but less is known (Yan et al., 2016). Prompt linguistic features are also known to influence EI test quaity, particularly item difficulty, but item discrimination or corpus-based, fine-grained meausres have rarely been incorporated into examining the contribution of prompt linguistic features. The current study addressed the research needs, using item response theory (IRT) and random forest modeling.

Data consisted of 9,348 oral responses to forty-eight items, including EI prompts, item scores, and rater comments, which were collected from 779 examinees of an L2 English EI test at Purdue Universtiy. First, the study explored the current and alternative EI scales / scoring methods that measure grammatical / semantic accuracy, focusing on optimal IRT-based measurement qualities (RQ1 through RQ4 in Phase I). Next, the project identified important prompt linguistic features that predict EI item difficulty and discrimination across different scales / scoring methods and proficiency, using multi-level modeling and random forest regression (RQ5 and RQ6 in Phase II). The main findings were (although not limited to): 1) collapsing exact repetition and paraphrase categories led to more optimal measurement (i.e., adequacy of item parameter values, category functioning, and model / item / person fit) (RQ1); there were fewer misfitting persons with lower proficiency and higher frequency of unexpected responses in the extreme categories (RQ2); the inconsistency of qualitatively distinguishing semantic errors and the wide range of grammatical accuracy in the minor error category contributed to misfit (RQ3); a quantity-based, 4-category ordinal scale outperformed quality-based or binary scales (RQ4); sentence length significantly explained item difficulty only, with small variance explained (RQ5); Corpus-based lexical measures and phrase-level syntactic complexity were important to predicting item difficulty, particularly for the higher ability level. The findings made implications for EI scale / item development in human and automatic scoring settings and L2 English proficiency development

CHAPTER 1. INTRODUCTION

Elicited imitation (EI) is an oral sentence repetition task that measures oral proficiency based on the degree to which examinees accurately repeat a given aural prompt (Underhill, 1987; Vinther, 2002). The construct of EI, including authenticity, has long been debated (Erlam & Akakura, 2016; Van Moere, 2012; Yan, Maeda, Lv, & Ginther, 2016). While criticism of EI for the lack of authenticity has existed, proponents of EI argue that accurate repetition is an outcome of reconstruction of the given prompt via internalized grammar rather than by rote memorization, because without reconstructing the aural prompt using their own linguistic knowledge (e.g., morphology, syntax, semantics, phonology, and phonetics), examinees cannot generate exact repetition (Jessop, Suzuki, & Tomita, 2007). Literature has provided evidence for EI as a measure of diverse aspects of language proficiency: implicit knowledge (Bowles, 2011; Ellis, 2005, 2009b; Erlam, 2006, 2009; Rebuschat, 2013; Serafini, 2013) or the efficacy of processing linguistic knowledge (Van Moere, 2012), and/or global oral proficiency (Christensen, Hendrickson, & Lonsdale, 2010; Cook, McGhee, & Lonsdale, 2011; Cox & Davies, 2012; Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008; Henning, 1983; Kahng & Otonya, 2021; Markman, Spilka, & Tucker, 1975; Naiman, 1974; Rebuschat & Mackey, 2013; Tracy-Ventura, McManus, Norris, & Ortega, 2014). Despite the divergence on the specific construct of EI, EI has been widely used as a reliable, efficient, and practical measure of L2 proficiency in research, standardized testing, and local / classroom assessment settings (Vinther, 2002; Yan et al., 2016; Zhou, 2012). Importantly, however, the effectiveness of EI depends on appropriate implementation of task features. Vinther (2002) highlighted scoring methods, prompt features (i.e., sentence length, grammatical features), and delayed repetition as key task features for valid measurement of L2 proficiency using EI. The current study focuses on two aspects of the task features, scoring and prompt features.

Selection of scoring methods largely impact of the effectiveness of EI (Vinther, 2002, Yan et al., 2016) because scales and scoring methods are central to reliability and validity of test scores (Fulcher, 1987; Knoch & Chapelle, 2018; Knoch, Deygers, & Khamboonruang, 2021; Weir, 2005). Therefore, it is crucial to examine how widely used EI scales and scoring methods influence reliability and validity of EI test scores. Despite the range of scoring methods used in EI (e.g., binary, ordinal, or interval scoring), information lacks regarding the impact of these scales and scoring methods on the effectiveness of EI. Noting the gap as well as the centrality, Yan et. al (2016) conducted a meta-analysis and reported outperformance of ordinal scales in distinguishing examinees of different L2 proficiency. Except for the theoretical or empirical synthesis (e.g., Vinther, 2002; Yan et. al, 2016), research on EI scoring is scarce. In particular, direct empirical comparison of different scoring methods has not been conducted within an individual study, which results in the lack of information on how the choice of scoring methods impacts specific and comprehensive qualities of EI test and items. Research on the impact of scoring methods would benefit item developers by providing strengths and weaknesses of different scoring options in relation with test and item performance. For example, useful characteristics of item and testing qualities can include (but not limited to) item difficulty, item discrimination, the adequacy of scale levels (e.g., differentiation of exact repetition from appropriate paraphrase), (mis)fits, coverage of proficiency levels, and reliability. In addition, along with the comparison of different scoring methods, details on the procedures of developing and/or revising EI scales and scoring methods based on the utilization of diverse sources (e.g., empirical scoring, rater feedback, L2 theories) would also be instrumental for future development of items, rubrics, and scales. The information would also offer an additional example of EI context to the discussion on the relationships between scales/scoring methods and construct validity of test scores for L2 testing researchers.

Prompt features are another decisive factor for the effectiveness of EI. Literature have reiterated the importance of prompt length for EI test sensitivity (Campfield, 2017; Miller, 1973, Perkins, Brutten, & Angelis, 1986, Yan et. al, 2016) as well as prompt linguistic features (Campfield, 2017; Graham, McGhee & Millard, 2010; Menyuk, 1971; Ortega, 2000). While the plethora of EI studies have already provided valuable information on the impact of prompt length and linguistic features, useful information can be further provided for EI test / item development, L2 proficiency research, and L2 instruction by (1) broadening the range of linguistic features, (2) comparing the impact of prompt length and linguistic features, (3) including important test qualities in addition to item difficulty, and (4) refining the conditions of examination.

First, the current EI literature lacks information on the predictability of prompt linguistic complexity/sophistication based on the fine-grained linguistic measures. With the recent development of automatic natural language processing and computational techniques, a wider range of fine-grained linguistic features have been examined in L2 speaking and writing tests, particularly from a usage-based, corpus linguistic approach (e.g., Biber, Gray, Poonpon, 2011;

Biber, Gray, & Staples, 2016; Crossley & Kyle, 2018; Kyle & Crossley, 2015; Kyle, Crossley, & Berger, 2018; LaFlair & Staples, 2017; Zhou, 2020). In addition, studies have examined syntactic and lexical features of prompt for item difficulty, but the comparison has been less investigated. These more comprehensive approaches of examining the impact of prompt linguistic features would be useful to enrich the guidelines for best practice of item development as well as provide additional insights into understanding L2 proficiency development and applying the results of EI test to L2 instruction.

Second, previous studies have extensively examined prompt length and linguistic features separately, or with only one or two linguistic components. Relatively few has investigated length and linguistic features of EI prompts with a comprehensive range. Acknowledging the practical benefits of using prompt length to adjust item difficulty, particularly for local and classroom settings—because sentence length is easy to measure—information on the relative effectiveness or importance of prompt length compared with linguistic features would practically benefit EI item developers. In addition, the comparison would add meaningful insights into EI research.

Third, although item discrimination is as an important item parameter as item difficulty for item and test quality, little is known about the impact of prompt length and linguistic features on item discrimination. The relationships would be useful for item development and EI testing research. For example, when item developers manipulate prompt features to adjust item difficulty, the information about the relationship aids the overall impact on item performance. Also, the information would broaden the understanding of item parameters of L2 EI and test in general.

Finally, examining factors that might influence the relationships between prompt features and item parameters would improve the applicability of the study results. Given the crucial role of scales and scoring methods, the impact of prompt features might vary depending on the selected scales and scores. Also, the test effectiveness or sensitivity is inherently related to the level of proficiency intended to distinguish. Therefore, it would be beneficial to incorporate scales / scoring methods and proficiency levels into researching the predictability of item parameters by prompt length and linguistic features should and item parameters.

Responding to these gaps, the current study first examined the impact of scales and scoring methods on a wide range of test and item measurement qualities using item response theory (IRT), including misfit analysis and the procedures of suggested revision of scales / scoring methods. Next, length and linguistic features of EI prompts were comprehensively examined in relation to their impact on item parameters within and across different scales /scoring methods and proficiency levels, using multi-level modeling (MLM) and random forest (RM) approaches.

CHAPTER 2. LITERATURE REVIEW

The theoretical foundations of the current dissertation were built on three aspects of EI: the construct, scoring, and prompt linguistic features. Beginning with the construct that EI measures, the discussion of the literature was centered on scales / scoring methods and prompt linguistic features of EI, situated within L2 assessment, acquisition, and corpus linguistics. Based on the centrality and needs about scoring and prompt features of EI that literature identified, six research questions were posed.

2.1 Construct Measured by Elicited Imitation

The construct of EI has been controversial among the L2 researchers over decades, receiving mixed evaluations in line with the alterations of main theoretical perspectives on L2 acquisition, teaching, and assessment (Yan et al., 2016). After the initial popularity as a measure of (grammatical) linguistic competence in the 1970's and 1980's (e.g., Henning, 1983; Markman et al., 1975; Naiman, 1974; Underhill, 1987), EI underwent construct validity-related criticism for inauthenticity (Hood & Lightbown, 1978; Hood & Schieffelin, 1978; Prutting, Gallagher & Mulac, 1975) and unclear or irrelevant construct representation (Fraser, Bellugi, & Brown, 1963; Gathercole & Baddeley, 1993; McDade, Simpson, & Lamb, 1982) as communicative teaching and learning evolved as the main approach. McDade et al. (1982) argued that examinees' working memory or short-term memory capacity is measured rather than language proficiency when examinees perform EI tasks or parrot what they listen to. In line with this view but with a greater focus on authenticity, Prutting et al. (1975) contended that the simple method of repeating aural cues did not seem to reflect the process of producing language for real-world communication which limits the construct representation to experimental settings.

Most recently, however, considerable interest in EI has been renewed, aligning with revitalized attention to information processing to assessing and understanding language proficiency. Researchers from a processing perspective have argued for EI as a measure of processing competence, while others have used EI as proxy for global oral language proficiency (Erlam & Akakura, 2016; Gass, 2018; Yan et al., 2016).

2.1.1 EI as a Proxy for L2 Oral Proficiency

EI has been extensively used as a proxy for global oral language proficiency for a wide range of L2s, including English (Bernstein, van Moere, & Cheng, 2010; Ortega, Iwashita, Norris, & Rabie, 2002). The contexts that EI covers are wide from large-scale and local standardized testing (e.g., van Moere, 2010; X. Li, 2020) to classroom assessment settings (e.g., Kahng & Otonya, 2021). Particularly related to testing approaches that benefit from automated scoring, EI, as a representative psycholinguistic task, has played foundational role (e.g., EI test in Phone Pass, Pearson, Duolingo English Test, and Essentials.

The traditional use of EI has been championed not only due to its (widely-known) high reliability and practicality in item development and test administration but also due to strong validity evidence (Erlam & Akakura, 2016; Gass, 2018; Yan et al., 2016). The effectiveness and use of EI is often examined in terms of the relationship between EI and another representation of the intended construct, a.k.a. criterion-related validity evidence. In the case of L2 English, Christensen et al. (2010) found a high correlation (r = 0.75) between EI test performance (N = 127) and the Speaking Language Assessment Test (SLAT). Similarly, Graham et al. (2008) found their EI test (N = 232) correlated with English oral proficiency interview (OPI) scores (r = 0.66) and informal placement interview scores (r = 0.64). In addition to speaking tasks, Okura and Lonsdale (2012) noted that EI scores (N = 40) were highly correlated to language center placement test results (r = 0.79), which included grammar, writing, and listening tasks. Cox and Davies (2012) further examined the relationships with specific sub-skills. Cox and Davies found EI (N = 179) was more highly correlated to listening scores (r = 0.74) than with grammar (r = 0.58) or reading scores (r = 0.60) while EI explained 47% of the variance in OPI scores. Yan et al. (2016) further unpacked comprehensive information on the performance of EI via a meta-analysis, which showed that EI can reliably differentiate examinees of high and low proficiency (Hedges's g = 1.34, SD = 0.13). Some studies further examined correlation between L2 proficiency measured by EI and prediction of performance intended by EI, such as scores of another proficiency test. For example, Cook et al. (2011) reported high prediction of actual OPI scores (r = 0.80 or higher) by EI results (N = 85), which was rated using an automatic speech recognition (ASR) engine.

Research in the setting of other L2s has also supported EI as a measure of global oral proficiency. Regarding L2 French, Tracy-Ventura et al. (2014) found EI performance (N = 29) correlated with oral narrative task scores (r = 0.67). Interestingly, Gaillard and Tremblay (2016) examined EI (N= 100) for speaking self- assessment, which resulted in 65% of variance explained. Similar, moderate to strong positive relationships between EI and oral proficiency or placement test results in L2 Spanish (Bowden, 2016), Korean (Kim, Tracy-Ventura, & Jung, 2016), and Chinese (Wu & Ortega, 2013). All these studies demonstrated high reliability, as well (α = .92 or higher).

2.1.2 EI as a Measure of Processing Competence (or Implicit Knowledge)

In addition to the traditional use of EI, that is, as a proxy for L2 global oral proficiency, L2 researchers have paid substantial attention to EI as a measure of implicit knowledge from a cognitive perspective (Erlam & Akakura, 2016; Gass, 2018; Yan et al., 2016). The interest in measuring implicit knowledge was coupled with the increased awareness of differentiation between implicit and explicit knowledge (Ellis, 2009a). Ellis (2009b) focused on the ontology of knowledge for language proficiency, learning, and teaching and differentiated implicit knowledge from explicit knowledge. Implicit knowledge is characterized by “subsymbolic, procedural, and unconscious,” as opposed to explicit knowledge or “analyzed knowledge” (p. 38). In other words, implicit knowledge cannot be clearly verbalized or explained. Pathways to obtain implicit knowledge are generally assumed to be implicit learning and teaching, which do not explicitly focus on formal aspects of language (e.g., linguistic terms or grammar rules). While explicit knowledge is argued primarily obtained by formal instruction, implicit learning, often referred to as acquisition, occurs during and/or as a result of continuous communication where language is not the end goal but rather a means, which is often the case of L1 contexts. Ellis stressed that, although both types of knowledge are important for language proficiency and contribute to linguistic competence, implicit knowledge is crucial for spontaneous language use without conscious planning. In other words, implicit knowledge is key to effortless language use and processing speed. These characteristics are the epistemology of automaticity or processing competence defined in the literature on language processing, for example, in dual-mode processing (Skehan, 1998), instance theory of automaticity (Logan, 1988), or language processing model (Levelt, 1989). It should be noted that the connection between implicit knowledge and processing competence is an important theoretical ground for the use of EI from the processing and psycholinguistic perspectives (Van Moere, 2012).

EI has been an important tool of measuring processing competence or the use of implicit knowledge, particularly given the overdue attention compared with other components of communicative competence (Van Moere, 2012), potential underrepresentation of linguistic competence (Ellis, 2009a), and yet greater challenge of the construct operationalization (Ellis, 2005, 2009b). Ellis' (2005, 2009b) studies are prominent earlier examples of operationalizing the measurement of implicit knowledge. In these studies, Ellis specified seven keys to measuring implicit knowledge: (1) intuition-based ('feel'-based) responses rather than rules, (2) time-pressured performance, (3) focus on meaning, (4) consistency in responses, (5) high certainty of correctness judgement on the responses, (6) non-involvement of metalanguage for rules, and (7) preference for learners with informal learning from earlier age. EI appropriately meets the criteria, along with oral narrative tasks and timed grammaticality judgement tasks (TGJT). Therefore, strong relationships among these tasks can serve as construct-related validity evidence for the use of EI as a measure of implicit knowledge and clarify the uncertainty of the construct and the use of EI.

Several studies have provided validity evidence for EI as a measure of implicit knowledge. Ellis (2005, 2009b) found that their L2 English EI test scores were strongly correlated with TGJT and oral narrative tasks than with the other tasks of the test battery intended to measure explicit knowledge, and that EI loaded onto the same construct along with the tasks for implicit knowledge in a factor analysis. Related to a standardized test, IELTS performance, Erlam (2006) demonstrated that the L2 English EI test (N= 115) was correlated with the listening section (r = 0.72) and the speaking section (r = 0.67) while the EI test was highly reliable (KDR20 reliability = .98). Later, Bowles (2011) conducted conceptual replication of Ellis (2005) with L2 Spanish speakers, although with a small sample (N = 30), and confirmed that EI loaded with TGJT and an oral narrative, with a high correlation with an oral narrative (r = 0.78). Further, Bowles found outperformance of heritage learners of Spanish over L2 Spanish speakers on EI. In addition, this pattern of correlations and factor loadings among the three implicit knowledge-based tasks was consistent when examining EI tests that focused on specific grammar points. Sarandi (2015) found that ratings of 27 L2 EI items that tested third person singular 's' of the English language (N = 50) were highly correlated with an oral narrative task (r = 0.73). Spada, Shiu, and Tomita (2015) also reported that scores of 21 L2 EI stimuli that tested English passive structures (N = 90) loaded along with TGJT with significant correlations.

Moving forward, other studies provided evidence of non-construct-irrelevance of EI regarding working memory. In Okura and Lonsdale's (2012) study, scores from the L2 English EI test of 60 items (N = 40) were found to have a low and non-significant correlation (r = 0.25) with a working memory test. The finding supported that EI is not merely simple verbatim activated by working or short-term memory, which serves as important construct-related validity evidence particularly given the controversy surrounding the relationship between EI and working memory.

The findings from these studies support the rationale behind the advocacy of EI that exact reproduction can be accomplished only when participants decode and interpret the given stimulus. To elaborate this view, Jessop et al. (2007) propose a set of three cognitive processes that reconstructive EI involves: 1) processing a prompt; 2) internally reconstructing the cue; and 3) reproducing the sentence. These cognitive processes that EI forces test takers to undergo are akin to Levelt's (1989) oral communication processing model that involves input and output processing (Van Moere, 2012). While pointing out the efficacy of processing, or automaticity as an important component of communicative competence, Van Moere also stresses that this language processing often occurs via repetition in real communication. In other words, EI is an authentic task. Regarding the counter-evidence against the inauthenticity of EI, Van Moere forefronts repetition as a fundamental aspect of conversation, which is frequently observed in the form of a summary of or uptake for a conversation partner's utterance.

Van Moere (2012) further explains the reconstructive nature of EI in relation to automaticity and lexical chunks, rather than application of (syntactic) rules, in oral communication. Van Moere highlights the lexical-based approach to understanding, assessing, enhancing automaticity by attending to the facilitative role of memory-based lexical chunks in the process of repeating language. Lexical chunks are often referred to as formulaic sequence / languages, or multiple words or phrases stored in and retrieved from language users' long-term memory (Pawley & Syder, 1983). Because formulaic sequences are accessed as if they were a single lexical chunk, the use of formulaic sequences are automatic and effortless without activating rule-governed language production. Thus, the greater repertoire of formulaic sequences an examinee has, the higher level of automaticity is demonstrated than language production based on syntactic rules and individual lexis, which is a laborious, effortful process (Pawley & Syder, 1983; van Moere, 2012). In this regard, examinees are able to successfully reconstruct the given EI stimuli that exceed the length of syllables storable in a short-term memory, only when they use the formulaic sequences they have already internalized in their long-term memory.

Collecting construct-related validity evidence for a langue test is an ongoing process, and the neurological processing of linguistic knowledge is far from being crystal clear; however, evidence in support of the use of EI as a measure of L2 proficiency is growing. In the next section, important consideration for a valid and effective use of EI test are discussed.

2.2 Factors that Influence the Construct Validity and Effectiveness of EI Tasks

The validity evidence collected throughout the literature may not apply unconditionally. Vinther (2002) emphasized four key factors to be considered to use EI tasks as a valid and effective measure of L2 proficiency, which are scoring methods, prompt length, grammatical features, and delayed repetition. The following subsections discussed the first three key task features, scoring and two prompt features, which were found to be significant moderators (Yan et al., 2016) and are the main interests in the current study.

2.2.1 Impact of EI Scales and Scoring Methods on EI Test and Item Quality

The selection of scales and scoring methods influences many aspects of test quality, for example, construct validity and reliability of (the uses and interpretations of) L2 test scores, and test sensitivity (Fulcher, 1987; Kane, 2013; Knoch & Chapelle, 2018; Knoch, Deygers, & Khamboonruang, 2021; Weir, 2005). The rationale behind this argument is that scoring functions as a tool that not only practically but also conceptually frame L2 test performance and the claims made by the performance (Kane, 2013; Knoch & Chapelle, 2018). In other words, scales and scoring rubrics are a practical and conceptual framework that assigns L2 performance to a symbol such as scores and/or category levels based on the theoretical groundings of a given test and practical applicability.

Impact of EI Scales and Scoring Methods

As in my L2 tests, scoring potentially influences how valid and reliable a given EI test is as a valid (and reliable) measure of L2 proficiency (Vinther, 2002), but the impact of scales and scoring methods has rarely been investigated, which results in far less information compared to research-based guideline provided regarding other decisive EI features (Yan et al., 2016). Yan et al.’s (2016) meta-analysis identified a binary scale (k = 24) as the most frequently used scale among the 58 studies examined, followed by ordinal scales (k =15) and interval scales (k = 15). Binary scales assess EI performance with only two levels, yes to correct repetition and no to incorrect repetition (e.g., Ellis, 2005; Erlam, 2006, 2009). Ordinal EI scales employ three or more category levels assigned to responses of different levels of repetition accuracy (e.g., Chaudron, Prior, & Kozok, 2005; Markman et al., 1975). Interval scales use the number or proportion of errors produced in responses (e.g., Graham et al, 2008; West, 2012) or automated measurement of specific features (e.g., Cook et al., 2011; Cox & Davies, 2012; Graham et al, 2008; Lonsdale & Christensen, 2011; Trofimovich & Baker, 2007).

Given the centrality of scoring methods and little information on the topic in the literature, Yan et al.’s (2016) meta-analysis filled a meaningful gap. The study found that scoring methods significantly moderated the distinguishability of EI in that an ordinal scoring approach distinguished examinees of different levels of L2 proficiency better than the other approaches (Hedges’s g = 1.61, SD = 0.08). In terms of individual studies rather than syntheses such as meta-analysis, investigation of scoring methods is limited with peripheral attention only. Only a few studies involved two or more EI scoring methods. Most of the studies were validation studies for automated scoring in EI, which focused on either correlation between human rating and automated scoring (e.g., Trofimovich & Baker, 2007; Graham et al., 2008) or feature selection (e.g., Lonsdale & Christensen, 2011), not on the impact of different scoring methods on test effectiveness or validity. For example, Graham et al. (2008) directly included two different scoring methods used for human ratings, a four-point ordinal scale versus frequency of the total number of correctly repeated syllables and compared the scores to evaluate the performance of automated scoring, rather than on the comparison of the two scoring methods, and little information was provided about the impact of each scoring.

Examination of EI Scales and Scoring Methods in the Current Study

Considering that the different types of scoring methods are available, the comparison of the EI scoring methods would add useful information to maximize the effectiveness of EI tests, but little is known about the impact of scoring methods on EI test quality from direct comparison in an individual study, that is on the same data with other variables controlled. The current study compared the current and revised scales and scoring methods to examine how the choice of scoring methods impacts specific and comprehensive qualities of EI test and items.

Particularly, the impact of different scales and scoring methods was assessed based on measurement information extracted using item response theory (IRT). The parameters examined were item difficulty, item discrimination, the adequacy of scale levels (e.g., differentiation of exact repetition from appropriate paraphrase), (mis)fits, coverage of proficiency levels, and reliability. The measurement information itself, in addition to the exploration of optimal scales and scoring methods, might be useful for the testing program and the field, given that the majority of analyses of EI scoring is based on classical testing theory (CTT) (e.g., X. Li, 2020) with a relatively few Rasch examples (e.g., Campfield, 2017; Perkins et al., 1986) .

In addition, little has been discussed about optimal scoring methods regarding the degree of imitation, for example scoring exact imitation compared with errorless paraphrases, either from a CTT or an IRT perspective. While exact imitation is scored higher than errorless paraphrasing based on theory, empirical evidence is neither clear nor sufficient to support the validity of this distinction. For the fair and valid use of EI scores, as well as reliable discrimination of examinees between high and intermediate proficiency, the independent use of the two categories calls for empirical support.

Detailed introduction of the scale and rubric evaluation and revision procedures would also contribute to greater understanding of scale and rubric development. Fulcher (1987) championed the advantages of scale development based on examinees' actual performance while Knoch, Deygers, and Khamboonruang (2021) noted utilization diverse sources available to aid the construct validity of scales and rubrics (e.g., rater feedback, L2 theories), including observed examinee performance. These studies informed the evaluation and revision of the current EI scales and scoring methods, which would be instrumental for future EI test development and research while adding a new testing context, EI, to the current discussion on scale and rubric development.

2.2.2 Impact of EI Prompt Features on EI Scores and Item Parameters

In addition to scoring, Vinther (2002) identified two prompt features, prompt length and prompt linguistic features, as potentially influential factors for valid EI tasks. Unlike scoring methods, the impact of these prompt features has been of central interest to L2 EI researchers and substantially discussed in the literature.

Impact of Sentence Length

Prompt length, measured by the number of syllables of a given prompt, is a well-known contributor to EI scores (Miller, 1973; Perkins et al., 1986; Bley-Vroman & Chaudron, 1994; Graham et al., 2010). Several studies documented medium to large variation of EI scores explained by sentence length, 45% in Kim et al. (2016) and higher in Graham et al. (2010) and in Wu and Ortega (2013), 73% and 74% of the score variance, respectively. While studies are congruent that sentence length contributes to EI item difficulty, different viewpoints exist in respect to what the predictability by length indicates for the construct of EI.

Some researchers see the contribution of length as an indication of influence of working memory, and thus, as potential construct irrelevance, while others the influence has theoretical grounding of language development. Miller (1956) argues that seven (± 2) segments are the maximum number of units (i.e., syllables, words) that can be retained in working memory. EI researchers suggest that EI prompt length should exceed this widely accepted information span, for example, seven syllables, to avoid the activation of working memory during performing EI tasks (Bley-Vermon & Chaudron, 1994; Vinther, 2002). Bley-Vermon and Chaudron further explained that examinees' internal grammar enables them to cluster syllables or words into larger units (e.g., phrases, clauses), which reduces the number of units and facilitates a long prompt processed in working memory. However, Fouly and Cziko (1985) claimed working memory affects repetition with longer prompts as well, in that increased prompt length beyond the length storable in working memory heightens the demand in working memory, which lowered the capability of accurate repetition. Based on this viewpoint, Sarandi (2015) interpreted a higher correlation (r = -0.74) between length and repetition of grammatical stimuli (i.e., correct form of third person singular -*s*) as reflection of the influence of working memory—the increased burden on working memory lessened the attention available for the grammar measured. Further, Sarandi found a smaller correlation between length and correction of ungrammatical stimuli (r = -0.28), with which Sarandi argued correcting ungrammatical prompts can be a more valid measure of implicit knowledge than repeating grammatical prompts.

On the other hand, Wu and Ortega (2013) regarded the high correlation (R = -0.86) as backed by L2 theories. Their claim was that examinees of higher proficiency were able to comprehend longer prompts, that is, more information, and reproduce them, while examinees of lower proficiency could not because the information required to processed was beyond their capacity and thus more challenging for them. What construct that the impact of sentence length on EI scores represent requires continued research, Klem et al. (2015) found no evidence that EI measures working memory, and Okura and Lonsdale (2012) documented a low and insignificant correlation between EI and working memory, which lends support for Wu and Ortega's (2013) perspective.

Another area calls for research is concerned with the impact of EI sentence length on item quality beyond the prediction of score variation, for example, item difficulty and item discrimination. Although prediction of scores provides substantial information on item difficulty, when the scores are not examinee-invariant or item-invariant, the generalizability of the findings is limited. Direct examination of prompt length in regard to IRT-based item difficulty can provide clearer and more generalizable picture.

In addition, information about the relationship between prompt length and time discrimination is scarce. The closet study would be Yan et al.'s (2016) meta-analysis which examined the moderating effect of sentence length on EI test sensitivity. Yan et al. found that differentiating length of EI prompts significantly increased EI test sensitivity to distinguishing examinees of different proficiency levels than using equal-length prompts (Hedges's g = 1.51, SD = 0.07). The finding aligns well with the studies that observed significant influence of prompt length on EI scores. The increased sensitivity of the meta-analysis also signals the potential of a meaningful relationship between sentence length and item differentiation although the coding of the sentence length (i.e., binary coding of variation or non-variation of prompt lengths) allows for limited information only. The examination of specific sentence length (e.g., number of syllables as a predictor) regarding item discrimination coupled with item difficulty would provide useful information for item development and insights into understanding the relationship in assessment and measurement contexts.

Impact of Prompt Linguistic Features

Along with prompt length, linguistic features of the EI prompts are known to be crucial for using EI as a valid and effective measure of L2 proficiency. Previous discussion of EI prompt linguistic features revolves around two topics: grammaticality and linguistic features.

Grammaticality Vinther (2002) listed grammatical structures of EI prompts, grammaticality and/or authenticity in particular, as important consideration for EI to function as a valid measure. However, Vinther's argument on grammaticality has not gained full support from empirical studies. In Erlam's (2006) study which employed both grammatical and ungrammatical prompts, L1 speakers' (n = 20) performance differed from L2 examinees (n = 95) in that L1 examinees corrected ungrammatical prompts more frequently (91%) than L2 examinees did (61%). However, Yan et al.'s (2016) meta-analysis found that the inclusion of ungrammatical EI prompts did not significantly moderate the test sensitivity. In line with Yan et al., Sarandi (2015) also reported that correlations of EI scores with oral narrative test scores were similar between grammatical (r = .62, p < .01) and ungrammatical prompts (r = .66, p < .01) in terms of third person singular -*s*. It is particularly interesting that grammaticality made a difference when the two types of prompts were correlated with sentence length.

Linguistic Features In addition to global L2 proficiency or implicit knowledge, EI has been used to measure one or more individual components of linguistic competence (e.g., lexical, syntactic, and/or morphological / morphosyntactic complexity / range / sophistication) or specific linguistic features. Yan et al. (2016) noticed syntactic and morpho-syntactic features were most commonly assessed (35 out of 43 studies). Examples of syntactic and morpho-syntactic features assessed via EI include articles (Akakura, 2012), *for* and verb form *to be* in Spanish (Fiori-Agoren, 2004), third person singular -*s* (Sarandi, 2015), past tense -*ed* (Ellis, Loewen, & Erlam, 2006), noun-adjective agreement of French (Erlam & Loewen, 2010), classifiers and perfective -*le* of Mandarin (S. Li, 2010), modals (Faqeih, 2012), wh-movement (Kim, 2012), passive (Spada et al., 2015), and so forth. Studies also found EI scores are related to lexical measures such as lexical frequency (Graham et al., 2010) or lexical range (Tracy-Ventura et al., 2014). Given linguistic competence is part of global proficiency or implicit knowledge required for comprehension, decoding, and reconstruction of EI prompts, the use of EI to assess specific linguistic competence or feature, and the correlations, is not unexpected. The connection emphasizes the role of linguistic features in controlling EI item difficulty. Indeed, Yan et al. (2016) found approximately half of the studies examined (36 out of 58) controlled syntactic features, and/or other lexical (9 studies), morphological (8 studies), and phonological features (6 studies).

Given the linguistic competence and features assessed and controlled in EI, no wonder item difficulty depends on linguistic features, such syntactic, and morphological/morpho-syntactic, and lexical features. Perkins et al. (1986) found derivational complexity is positively associated with Rasch-based item difficulty of 18 EI items taken by 50 L2 examinees. Perkins et al. also revealed that most difficult features for the examinees were processing adverbials, compounded and reduced clauses, and non-finite verbal phrases such as gerunds, past / progressive participles, and infinitives. Diverse lexical measures were also examined in respect to item difficulty. Graham et al. (2010) looked into 60 EI prompts in respect to lexical difficulty, and found lexical frequency (i.e., seven brackets of frequency levels on Kilgarriff's lemmatized list) and lexical density (i.e., ratio of content words to the total number of words) contributed to 8% and 2% of the variance in EI scores, while morphological complexity (i.e., morpheme-based word length) did not significantly predict the EI scores. Meanwhile, Campfield (2017) examined Rasch-based item difficulty and CCT-based average scores of 40 prompts in relation to lexical complexity, as measured by lexical density, function word density (i.e., ratio of function words to the total number of words), and morphological complexity, found none of the measures significant for item difficulty.

Linguistic Features and Sentence Length It is interesting that, in all the three studies (i.e., Campfield, 2017; Graham et al., 2010; Perkins et al., 1986), prompt length outperformed linguistic features in explaining the variability in item difficulty and/or scores of EI. Perkins et al. found the number of words (r = 0.88) and syllables (0.87) were more highly correlated with item difficulty (logits) than (morpho-)syntactic features. The number of syllables explained far more than lexical difficulty, which accounted for 73% of the score variation in Graham et al.'s study, and 31% in Campfield's study. Note that the range of the prompt lengths was fairly wide and included very short length (i.e., shorter than seven syllables): three to eighteen in Perkins et al, four to nineteen in Graham et al., and four to thirteen in Campfield. The examined outperformance of prompt length might not be consistent when items shorter than Miller's (1956) magic number (i.e., approximately seven), or the range is smaller. Nevertheless, it can be concluded that sentence length is potentially more important than (morph-)syntactic or lexical features of EI prompts in predicting EI item difficulty.

One important topic less investigated regarding prompt linguistic features and item difficulty is the comprehensive analysis of linguistic features. Examining lexical, syntactic, and morpho-syntactic features as well as sentence length would deepen the understanding of linguistic contributors to item difficulty in the context of EI and psycholinguistic assessment broadly.

Currently, few studies have tacked the comprehensive analysis. Some studies considered a comprehensive list of linguistic features for automatic scoring (Graham et al., 2008; Lonsdale & Christensen, 2010) or item development (Christensen et al., 2010) but item difficulty for each syntactic, lexical, and morphological variable was neither focused nor specified. Hendrickson, Aitken, McGhee, and Johnson (2010) is the only study, to my best knowledge, that considered all three linguistic dimensions at the feature level as well as sentence length with a clear focus on their contributions to EI item difficulty. Hendrickson et al. also found syllable count as the most important predictor in their best-performing model of the step-wise regression, followed by tense and aspect features. The outperformance aligned with previous studies (e.g., Campfield, 2017; Graham et al., 2010; Perkins et al., 1986). Interestingly, Hendrickson et al. further constructed models by syllable bands (i.e., seven, eight, and nine syllable-long prompts) and noted that the contributions and ranks of the variables varied across the models, which means potential interactions between linguistic features and prompt length, and broadly with proficiency levels. The findings invite studies that examine the comprehensive linguistic features with a wider range of sentence length and include proficiency levels and interactions.

Corpus-Based Measures and Formulaic Sequence The relationships between prompt linguistic features and item difficulty / scores of EI have been examined using traditional measures of lexical and syntactic features (e.g., Campfield, 2017; Graham et al., 2010; Perkins et al., 1986; Wu & Ortega, 2013). Recently, however, in L2 studies on linguistic analysis of other communicative skill performance tasks (e.g., speaking and writing free responses) the use of, so called, corpus-based measures has been on steady increase (e.g., Biber et al., 2011, 2016; Crossley & Kyle, 2018; Kyle & Crossley, 2015; Kyle et al., 2018; LaFlair & Staples, 2017). These corpus-based measures have some features distinguishing from traditional measures. Traditional measures are mostly length-based, and heavily rely on T-units, clauses, and sentence (e.g., Lu's (2010) Syntactic Complexity Analyzer (SCA), Lu's (2012) Lexical Complexity Analyzer (LCA)). Corpus-based measures consider phrases and n-grams as well as clausal level units, while the strength of associations and frequency in the reference corpus are decisive factors of the measurement of indices rather than (or in addition to) simple counts (Kyle, 2016, Kyle et al., 2018).

Indices from two open-source corpus-based analyzers can be good examples: the Tool for the Automatic Assessment of Syntactic Sophistication and Complexity (TAASSC) (Kyle, 2016) and the Automatic Analysis of Lexical Sophistication (TAALES 2.0) (Kyle et al., 2018). For syntactic analysis, TAASSC provides a comprehensive list of fine-grained complexity features both at clausal and phrasal levels (e.g., dependents counts per nominal, occurrence of particular dependents) as well as syntactic sophistication measures such as verb argument constructions (VAC) based on diverse sub-corpora of Corpus of Contemporary American English (COCA), in addition to traditional measures from SCA (i.e., fifteen indices based on T-units, clauses, dependent clauses, and verb / noun / coordinate phrases). Similarly, lexical indices extracted from TAALES 2.0, include frequency, range, and strength of association of n-grams, contextual distinctiveness, word recognition norms, semantic network. Classical measures extractable from LCA (e.g., lexical density, lexical sophistication, and lexical variation) are also included.

These corpus-based measures are not necessarily better indicators of linguistic complexity / diversity / sophistication than traditional measures, although it is possible to perform better in some data context. Rather, the fine-grained indices can measure different aspects of linguistic complexity from a usage-based approach, thus different implications would be made for L2 test development, research, and instruction (Biber et al., 2011, 2016; Crossley & Kyle, 2018; Kyle & Crossley, 2015; Kyle et al., 2018; LaFlair & Staples, 2017). However, the corpus-based measures have rarely used for the analysis of EI prompt linguistic features for item difficulty, which results in imbalance in information in general, and less from a usage-based perspective. Formulaic sequence is fixed or frequently-cooccurring n-grams used in communication. The role of formulaic or lexical structural stem in EI performance has been well supported both theoretically (e.g., Van Moere, 2012) and empirically (e.g., Yan, 2015). Given the connection between n-gram-based measures and formulaic sequences, that measurement from corpus-based indices can improve the understanding of item difficulty and L2 performance in EI and psychological testing contexts.

Examination of Important Prompt Features in the Current Study

The review of literature on prompt features indicated the central role of prompt length and linguistic features in item difficulty, as well as the potential benefit of more comprehensive linguistic analysis, including interactions, using both traditional and corpus-based measures. The current study examined prompt length, classical and fine-grained features of three linguistic components (lexis, syntax, and morphology), including their interactions, for item difficulty. Particularly, in addition to item difficulty, item discrimination, which is another important item quality, was predicted separately, while the relationships were examined across different proficiency levels, scales /scoring methods. To address the statistical issues that the complexity and quantity of variables raise, random forest regression, a machine learning method, was used.

2.3 Research Questions

Based on the literature review, six research questions (RQs) were posed in search of optimal EI rating scales and scoring methods for EI measurement qualities, focusing on assessment of item parameters, (mis)fit, and category adequacy (RQ1 to RQ4, Phase I), and identify important prompt features to predicting item difficulty and discrimination (RQ5 and RQ6, Phase II). The six main RQs, including the sub-RQs for step-by-step investigation of each RQ, are provided as follows:

RQ1. What is the optimal number of EI response categories for EI measurement qualities, particularly regarding the independent use of exact repetition and appropriate paraphrase?

1.1 Does a 4-category EI scale, which collapses the lowest category (i.e., no, incomprehensible, or irrelevant response) into its adjacent category (i.e., response with major grammatical errors or meaning difference), more optimally measure accuracy than the current 5-category scale?

1.2 Does an IRT model on the 4-category EI scale support the higher ordering of exact repetition than errorless paraphrase and the use of the two separate categories?

1.3 What item characteristics are associated with the appropriateness of using the two separate categories?

RQ2. Which examinee ability and response category levels have the largest number of person and item misfits?

2.1 Where did the person misfits of the 4- and 3-category EI accuracy scales most occurred in terms of person ability and response category levels?

2.2 Where did the item misfits of the 4- and 3-category EI accuracy scales most occurred in terms of person ability and response category levels?

RQ3. What are the potential sources of person / item misfits and category inadequacy?

3.1 What are the potential sources of the person misfits on the 4- and 3-category EI accuracy scales in relation to rating grammatical errors and semantic deviations?

3.2 What are the potential sources of the item misfits on the 4- and 3-category EI accuracy scales in relation to rating grammatical errors and semantic deviations?

3.3 Do the issues with rating grammatical errors and semantic deviations observed in misfitting items and examinees exist in item responses with unexpected scores of non-misfitting items or examinees?

3.4 What are the potential source of the inadequacy of using the paraphrase category?

3.5 What guidelines for item development, scale/rubric revision, and rater training do the qualitative analysis provide to minimize misfits and increase the adequacy of using the paraphrase category?

RQ4. What is the optimal EI scale / scoring method for measuring L2 semantic and grammatical accuracy?

4.1 What are the alternative EI accuracy scales and rubrics that address the issues of rating criteria in semantic and grammatical judgement?

4.2 Do the quantity-based, alternative EI ordinal scales and rubrics perform better than the quality-based, original EI ordinal scales and rubrics?

4.3 Does the best fitting ordinal scale perform better than binary scale?

RQ5. What are the relationships between EI prompt length and EI item parameters (i.e., item difficulty and item discrimination) across scales / scoring methods?

5.1 To what extent, does prompt length (i.e., number of syllables, number of words) impact item difficulty across eleven different scales / scoring methods?

5.2 (a) To what extent, does prompt length (i.e., number of syllables) impact item discrimination across ten different ordinal scales / scoring methods? (b) Which characteristics of scale modification are associated with the relationship?

RQ6. Which prompt linguistic features are most important to predicting EI item parameters (i.e., item difficulty and item discrimination) across response category levels and scales / scoring methods? Are there any important interactions?

6.1 (a) Which linguistic features of EI prompts are important to predicting EI item difficulty across the category levels and different scales? (b) Which features interact with category levels of difficulty and/or scales / scoring methods?

6.2 (a) Which linguistic features of EI prompts are important to predicting EI item discrimination across the different scales? (b) Which features interact with scales/scoring methods?

CHAPTER 3. METHODS

3.1 Instrument: ACE-In Elicited Imitation

3.1.1 Context: the Assessment of College English-International (ACE-In) in the Purdue Language Culture Exchange (PLaCE) Program

The instrument of the study is an EI test called *Listen and Repeat*, which is a subsection of the Assessment of College English-International (ACE-In). The ACE-In is a locally-developed, post-entry L2 English exam for international students admitted to Purdue University. The ACE-In was developed by the Purdue Language Culture Exchange (PlaCE), which provides language and cultural support via two English courses (i.e., ENGL 110, ENGL111) for international L2 undergraduate students with TOEFL iBT total scores between 80 and 100 (or an IELTS band scores of 6.5 to 7.0). All students who enroll in PLaCE courses take the ACE In as part of program evaluation, as well as international L2 English students who do not have a standardized English test scores on admission who take the ACE-In for placement. Towards the end of the course, students take a post test. The ACE-In (and the EI test) are used:

- to inform the baseline for L2 English instruction
- to diagnose examinees' language needs for their English learning
- to establish L2 English sub-skill profiles of international L2 English students
- to examine L2 English development (trajectory) via the courses
- to evaluate the language/cultural program's achievement
- to provide advisory information to inform placement and exemption decisions

The ACE-In consists of three modules and four tasks: a cloze elide task and an EI task for Module 1, independent and integrated speaking tasks for Module 2, and a timed essay writing task for Module 3. The test is Internet-based and administered in a university's lab via a standardized procedure. The present study is concerned with the EI task only, and the details are provided below.

3.1.2 ACE-In Elicited Imitation

The ACE-In EI has four forms, with each consisting of twelve items. The four forms developed to be comparable. The number of medium (15 to 17 syllables) and long (19 to 21 syllables) prompts are fixed based on the number of syllables, the range of lexical difficulty is specified, and the topics are related to campus life by and across forms.

During the EI test, one of the four forms is randomly assigned to examinees. Examinees take the test by listening to a prompt, clicking on a related word, and then repeating the prompt. The step of clicking on a word results in delayed repetition. The intentionally inserted distractor is to prevent mechanical verbatim. Examinees' responses are automatically recorded and uploaded onto the web-based ACE-In platform.

Scoring is based on a five-category ordinal scale, which classifies the different levels of repetition in terms of semantic and grammatical accuracy. Table 3.1 presents the rating scale, which ranges from omission (i.e., score 0) to exact repetition (i.e., score 4). Two randomly selected trained raters score each item of one test. The total scores, not item scores, between the two raters are compared. Differences of five points or larger are considered discrepant scores and are assigned to a third rater. The average scores of two agreed ratings, which is different from each other smaller than 5 points, are assigned as a final score.

Table 3.1 Scoring rubric of the ACE-In EI

Category	Score	Description
Exact repetition	4	Repeating the prompt exactly word for word*
Appropriate paraphrase	3	Paraphrasing the prompt with no grammatical errors and same meaning
Minor deviation	2	Paraphrasing the prompt without distorting meaning (i.e., keeping the same main idea) and/or with minor grammatical errors
Major deviation	1	Paraphrasing the prompt with distorted meaning (i.e., changing the same main idea) and/or with major grammatical errors
Omission	0	No response or response with only a few words that does not independently make sense

Notes: * contracted forms are not penalized

Expanding on the current practice, the present study conducted item-level examination from an IRT perspective. During the preliminary analyses, CTT-based information about rater and item performance were examined. Table 3.2 provides inter-rater reliability (i.e., Spearman correlation) among seventeen raters based on their initial ratings. Overall, inter-rater reliability was .80 with the range from .66 to .94. Seven pairs of raters out of the total of 82 pairs showed reliability lower than .70 while the performance of five pairs was very high (r = .90 or higher). The overall rater reliability (r = .80) was high but improvement for more reliable scores is still beneficial for accurate information used for exemption decisions and instructional purposes. In addition, the range is fairly wide, which invites to investigate the sources of rater inconsistency on item level.

Table 3.2 Inter-Rater Reliability Among Seventeen Raters (Spearman Correlation)

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17
R1	1																
R2	0.75	1															
R3	-	-	1														
R4	0.83	0.82	0.90	1													
R5	0.67	0.71	-	0.85	1												
R6	0.77	0.78	-	0.78	0.77	1											
R7	0.79	0.68	-	0.82	0.69	0.69	1										
R8	0.66	0.76	-	0.78	0.83	0.77	0.88	1									
R9	0.76	0.87	-	0.79	0.74	0.74	0.73	0.84	1								
R10	0.81	0.74	-	0.86	0.78	0.87	0.87	0.74	0.87	1							
R11	0.83	0.78	-	0.91	0.82	0.79	0.82	0.71	0.82	0.76	1						
R12	0.69	0.76	-	0.78	0.76	0.80	0.69	0.71	0.77	0.82	0.84	1					
R13	0.84	0.85	-	0.87	0.85	0.78	0.85	0.85	0.83	0.90	0.81	0.82	1				
R14	0.84	0.87	-	0.76	0.80	0.84	0.90	0.82	0.87	0.74	0.94	0.76	0.83	1			
R15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1		
R16	-	-	-	-	-	-	-	-	-	-	0.73	0.89	-	-	-	1	
R17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.82	-	1

Internal consistency was also examined to evaluate item performance, using five indices: inter-item correlation, item-total correlation, Cronbach's alpha, split-half reliability (including adjusted reliability using the Spearman–Brown prophecy formula), and composite reliability. Table 3.3 presents the values of the five internal consistency measures for the forty-eight EI items. Overall, all the indices suggested good internal consistency. Regarding the initial ratings, both on test- and item levels, inter-item consistency showed no items with a too low or high correlation (i.e., r < .15 or r > .50), which means that items are distinct but correlated to appropriate degree. Item-total correlations were good, as well, with the minimum .51 (Item 48), which suggests good discrimination. Cronbach’s Alpha values also indicated high internal consistency with the range from 0.85 to .87. Split-half reliability was acceptable and expected to be increased when the test is lengthened from the adjusted values. Lastly, composite reliability, which is also known as construct reliability, was high with the range between .86 to .90. The high values or shared variance suggest that the items measure the same construct. Note that, however, these CTT-based internal consistency and composite reliability support the use of the EI items for as a measure of the intended construct, L2 oral English proficiency, in general, but do not give similar information on the category level. IRT-based analyses from the current project would provide further information to understand and refine the ordinal EI scale currently being used.

Table 3.3 Internal Consistency of Forty-Eight EI Items

a. Overall Internal Consistency

Form (Items)	Inter-item Correlation		Item-Total Correlation		Cronbach's Alpha		Split-half reliability (adjusted)		Composite reliability	
	Initial	Final	Initial	Final	Initial	Final	Initial	Final	Initial	Final
F1 (Q1 to Q12)	0.36	0.37	0.64	0.65	0.87	0.88	0.77 (0.87)	0.78 (0.88)	0.87	0.88
F2 (Q13 to Q24)	0.41	0.41	0.68	0.68	0.89	0.89	0.78 (0.88)	0.81 (0.89)	0.89	0.89
F3 (Q25 to Q36)	0.42	0.44	0 69	0.70	0.90	0.90	0.78 (0.88)	0.80 (0.89)	0.90	0.91
F4 (Q37 to Q48)	0.34	0.35	0.62	0.63	0.86	0.86	0.70 (0.82)	0.68 (0.81)	0.86	0.86
Average	0.38	0.39	0.66	0.67	0.86	0.87	0.76 (0.86)	0.77 (0.87)	0.88	0.89

b. Item-Level Internal Consistency

Form	Item	Inter-Item Correlation		Item-Total Correlation		Cronbach’s Alpha	
		Initial Scores	Final Scores	Initial Scores	Final Scores	Initial Scores	Final Scores
Form 1	Q 1	0.39	0.40	0.68	0.67	0.85	0.87
	Q 2	0.40	0.42	0.69	0.71	0.85	0.86
	Q 3	0.32	0.37	0.61	0.65	0.86	0.87
	Q 4	0.35	0.39	0.64	0.68	0.86	0.87
	Q 5	0.36	0.39	0.66	0.69	0.86	0.87
	Q 6	0.31	0.32	0.57	0.58	0.86	0.87
	Q 7	0.34	0.33	0.63	0.61	0.86	0.87
	Q 8	0.38	0.41	0.69	0.71	0.86	0.86

Table 3.3 continued

	Q 9	0.41	0.42	0.70	0.70	0.85	0.86
	Q 10	0.36	0.38	0.64	0.64	0.86	0.87
	Q 11	0.28	0.26	0.50	0.48	0.87	0.88
	Q 12	0.37	0.40	0.65	0.69	0.86	0.86
Form2	Q 13	0.45	0.45	0.73	0.73	0.85	0.87
	Q 14	0.38	0.37	0.63	0.62	0.85	0.86
	Q 15	0.39	0.41	0.65	0.68	0.86	0.87
	Q 16	0.43	0.42	0.71	0.70	0.86	0.87
	Q 17	0.44	0.43	0.73	0.71	0.86	0.87
	Q 18	0.40	0.39	0.67	0.66	0.86	0.87
	Q 19	0.40	0.40	0.68	0.68	0.86	0.87
	Q 20	0.44	0.45	0.74	0.75	0.86	0.86
	Q 21	0.38	0.37	0.62	0.61	0.85	0.86
	Q 22	0.36	0.38	0.60	0.62	0.86	0.87
	Q 23	0.37	0.36	0.61	0.59	0.87	0.88
	Q 24	0.46	0.47	0.75	0.76	0.86	0.86
Form3	Q 25	0.49	0.50	0.78	0.77	0.85	0.87
	Q 26	0.44	0.46	0.71	0.73	0.85	0.86
	Q 27	0.39	0.40	0.64	0.64	0.86	0.87
	Q 28	0.36	0.39	0.58	0.61	0.86	0.87
	Q 29	0.47	0.48	0.76	0.76	0.86	0.87
	Q 30	0.35	0.38	0.61	0.65	0.86	0.87
	Q 31	0.40	0.41	0.66	0.66	0.86	0.87
	Q 32	0.38	0.40	0.64	0.65	0.86	0.86
	Q 33	0.46	0.48	0.74	0.75	0.85	0.86
	Q 34	0.48	0.48	0.74	0.74	0.86	0.87
	Q 35	0.49	0.50	0.78	0.78	0.87	0.88
	Q 36	0.36	0.39	0.59	0.62	0.86	0.86
Form4	Q 37	0.30	0.29	0.58	0.55	0.85	0.87
	Q 38	0.38	0.37	0.69	0.68	0.85	0.86
	Q 39	0.31	0.33	0.60	0.61	0.86	0.87
	Q 40	0.31	0.34	0.57	0.59	0.86	0.87
	Q 41	0.37	0.36	0.70	0.68	0.86	0.87
	Q 42	0.38	0.38	0.70	0.69	0.86	0.87
	Q 43	0.33	0.35	0.62	0.64	0.86	0.87
	Q 44	0.31	0.33	0.60	0.62	0.86	0.86
	Q 45	0.34	0.34	0.62	0.61	0.85	0.86
	Q 46	0.38	0.39	0.68	0.69	0.86	0.87
	Q 47	0.35	0.37	0.64	0.66	0.87	0.88
	Q 48	0.27	0.30	0.51	0.56	0.86	0.86

Notes: Initial scores – the mean of the two initial scores rated by two raters; final scores – agreed scores between the two initial ratings or adjudicated scores by a third rater

3.2 Sample

The data for the current study were collected from 779 examinees who took an EI test as part of the ACE-In from Spring 2017 to Fall 2019, which resulted in the total number of 9,348 item scores across the EI four forms. The number of examinees was not equal due to the convenient sampling, but not drastically different. As shown in Table 3.4, the four forms appeared comparable in terms of the means and medians of the total scores although the minimum and maximum values varied to some degree.

Table 3.4 Data Collected for the Current Study

Form	Number of examinees	Number of item scores	Total score (out of 48)		Range	
			Mean	Median	Minimum	Maximum
1	193	2,316	25.24	24	11	42
2	202	2,424	25.39	25	8	47
3	204	2,448	26.32	24	8	45
4	180	2,160	25.91	25	13	43
Total	779	9,348	25.71	25	8	47

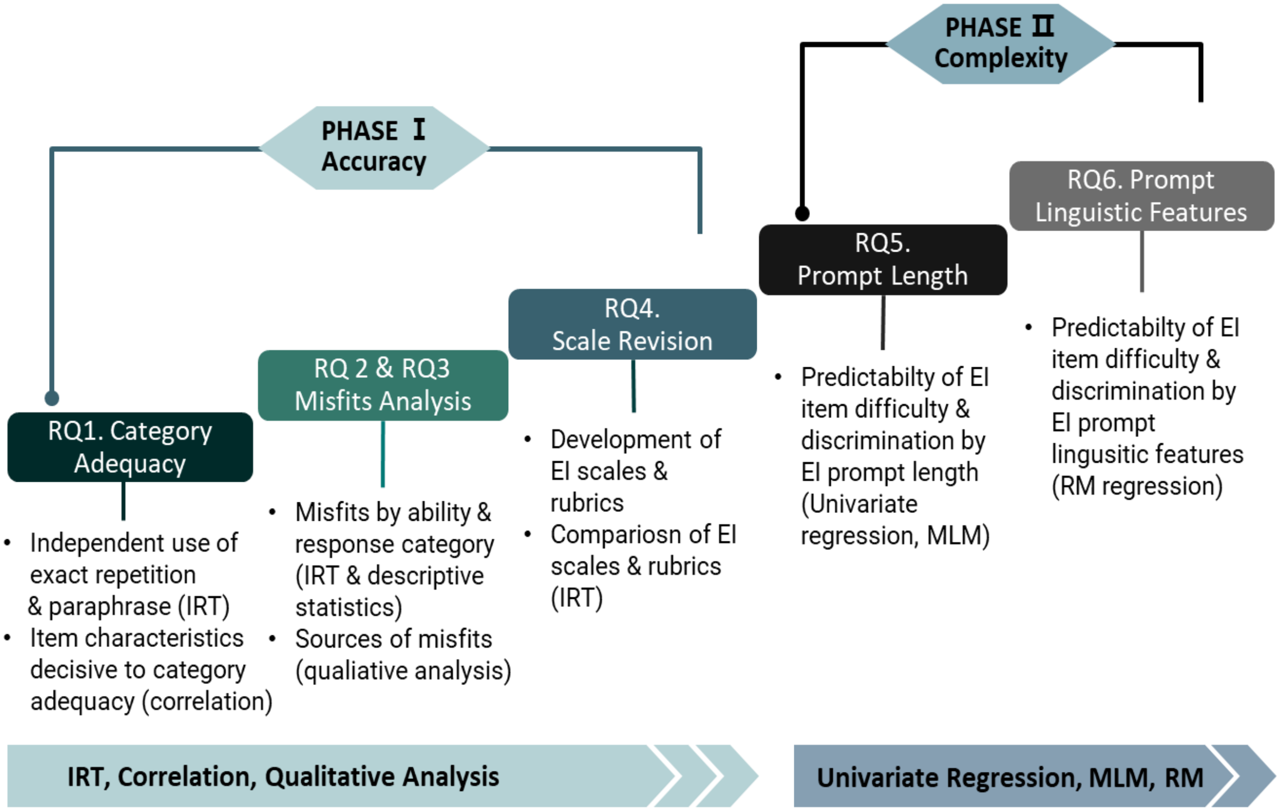
The examinees were diverse in their L1s and nationalities, coming from 68 different countries speaking 41 different languages. The most common L1 was Mandarin, accounting for 47.31% (n = 378), followed by Spanish (n = 74, 9.50%), Korean (n = 65, 8.34%), and Hindi (n = 60, 7.70%). The most frequent nationality was also China (n = 348, 44.67%), which far outnumbered the second and the third most frequent nationalities, India (n = 94, 12.07%) and South Korea (n = 65, 8.34%). Approximately two thirds of the examinees were male (n = 486, 62.39%). The majority of the examinees were enrolled in STEM programs.

3.3 Main Methodological Approaches

The current study is a two-phased project. Phase I was concerned with EI as a measure of semantic/grammatical accuracy by exploring scales/scoring methods for optimal measurement. Expanding on the results from the first phase, Phase II investigated the predictability of EI item difficulty and discrimination parameters by EI prompt features. The project flow and three main methods are presented in this section.

3.3.1 Flow of the Current Project with Main Methodological Approaches

Figure 3.1 illustrates the flow of the current project. The main topic of the six RQs and corresponding analysis methods are presented.



PHASE I
Accuracy
PHASE II
Complexity
RQ1. Category Adequacy
RQ 2 & RQ3 Misfits Analysis
RQ4. Scale Revision
RQ5. Prompt Length
RQ6. Prompt Linguistic Features
• Independent use of exact repetition & paraphrase (IRT)
• Item characteristics decisive to category adequacy (correlation)
• Misfits by ability & response category (IRT & descriptive statistics)
• Sources of misfits (qualiative analysis)
• Development of EI scales & rubrics
• Compariosn of EI scales & rubrics (IRT)
• Predictabilty of EI item difficulty & discrimination by EI prompt length (Univariate regression, MLM)
• Predictabilty of EI item difficulty & discrimination by EI prompt lingusitic features (RM regression)
IRT, Correlation, Qualitative Analysis
Univariate Regression, MLM, RM

Figure 3.1 A Flowchart of the Current Project with Main Methodological Approaches

The exploration of optimal scales / scoring methods (Phase I) began with investigating the optimal number of EI categories for psychometrically adequate measurement (RQ1). Using item response theory (IRT), I examined the technical adequacy of the use of the two highest scale categories, *exact repetition* and (appropriate) *paraphrase*. Next, among the IRT-based measurement statistics, item and person misfits were further examined in relation to examinee ability and response category levels, using IRT and descriptive statistics (RQ2), followed by qualitative analysis of the sources of misfits (RQ3). Finally, revised scales / scoring methods were developed and proposed, IRT-based measurement statistics of which were compared (RQ4).

Moving on to Phase II, the relationship between accuracy (measurement performance) and EI prompt features were examined, using the IRT-based item difficulty and item discrimination values across different EI scales and scoring methods examined in Phase I. First, item difficulty was regressed onto sentence length using univariate regression and then the association between item discrimination and sentence length was examined using a multi-level modeling (MLM) approach (RQ5). Next, important prompt linguistic features and their interactions for predicting the two item parameters were identified based on random forest (RM) regression (RQ5). The two main statistical methods used for the current projects are introduced, with applications to the current study, in the following sections.

3.3.2 Item Response Theory (IRT)

IRT is the main analytic framework for Phase I. IRT is a modern test theory that allows modeling the relationship between item responses and their underlying construct based on a non-linear monotonic function that associates examinees' ability levels on a latent trait with the probability of a particular response to a given item (de Ayala, 2013; Embretson & Reise, 2000). IRT often is often used as an alternative to Classical Test Theory (CTT) due to several advantages, such as the rich information on the individual item performance, detailed measurement precision across different levels (rather than a single estimate in CTT), and item-independent scores and sample-independent item parameters.

IRT models allow for fitting a range of outcome types, for example, one or two-parameter logistic models for binary dependent variables (Lord, 1980), the Generalized Partial Credit Model (GPCM, Muraki, 1992) or the Graded Response Model (GRM, Samejima 1969) for polytomous responses, and the Continuous Response Model (Samejima, 1973) for continuous outcomes. For the current project, unidimensional GRM was selected over GPCM, another model for ordinal responses, because GPCM penalizes responses with a greater number of categories. The penalization rare occurs in GRM because GRM considers items as a series of k-1 dichotomous items, where k means the total number of response categories. Rasch modeling was not considered, although checked for the appropriateness of GRM, because item discrimination of individual items was of central interest.

IRT generates several indices useful for evaluating measurement qualities on the items test levels. The IRT-based measurement statistics that the current study employed are presented in detail in Table 3.5. For the GRM modeling, the *mirt* package in the *R* programing language environment (Chalmers, 2012) was used based on full-information maximum likelihood estimation and an expectation-maximization (EM) algorithm.

Table 3.5 IRT-Based Measurement Statistics Used in the Current Study (de Ayala, 2013)

Statistics	Level	Definition and Usage
Item difficulty (*b*)	Item (category)	The location where (each category of) a given item functions on the scale. The lower values indicate easier categories / items.
Item discrimination (*a*)	Item	The degree to which a given item differentiates examinees of different ability levels. The lower values indicate lower discrimination.
Category characteristic curve (CCC)	Item (category)	A graphical description of the mathematical relationships between examinee's ability (ϴ) or underlying trait level and its responses to items on a scale. The plot for the current study has multiple curves, with each describing the probability of endorsing each response category (except for the lowest). CCC is used to evaluate category adequacy.
Information	Item & test	The amount of information in IRT is the measurement precision obtained by a given item or test on the scale that measures the construct, conditioned at ability levels. More highly discriminating items have greater information, which means greater precision.
Standard errors	Item (category)	Precision of estimating each item parameter at categorical and item level. Inversely related to information.
IRT reliability	Test	A single measure of the IRT marginal reliability. Higher values are higher reliability.

In IRT, the fit can be assessed at three levels: items, examinees, and test levels. For model fit assessment, general method of assessing factor analysis can be used. The model fit indices that the current project examined are:

- M_2 limited information goodness-of-fit statistic (Maydeu-Olivares & Joe, 2006)
- the Tucker-Lewis index (TLI, Bentler, 1990)
- the Comparative Fit Index (CFI, Bentler, 1990)
- the Root Mean Square Error of Approximation (RMSEA, Steiger, 1990)
- the Standardized Root Mean Square Residual (SRMR)

- the Akaike's information criterion (AIC, Bozdogan, 1987): −2log-likelihood plus twice the number of parameters
- the Bayesian information criterion (BIC; Schwarz, 1978): −2log-likelihood plus the logarithm of the sample size times the number of parameters

The model was considered good when RMSEA is 0.06 or smaller—while 0.05 or smaller is ideal with M_2 values (Maydeu-Olivares, 2015)—SRMR of 0.08 or smaller, and CIF and TLI of 0.95 or larger (Hu & Bentler, 1999). Smaller M_2 fit indices, AIC and BIC, were considered the preferred model. Noting the high sensitivity of M_2 values to even small misfits, associated *p*-values were considered (Toland, 2014), where a larger *p*-value is a better fit, and small values (p <0.05) are flagged. The criteria on the size of meaningful difference in AIC and BIC are not fixed, presumably, a BIC difference of 10 was a cut-off, which equals a Bayes factor of 150 (Raftery, 1995). In addition, the total amount of variance explained by the model and overall and individual factor loadings from an exploratory factor analysis (EFA) were examined, for all of which the larger values or loadings, the better fit. The information from EFA demonstrates overall model fit of the EFA model, discrimination, and dimensionality because individual factor loadings correspond to item discrimination.

IRT-based item and person fits are evaluated based on the deviations between predicted and empirical scores. Item fit was assessed using the $S\text{-}X^2$ statistic (Orlando & Thissen, 2000), which is a widely used approach for non-Rasch IRT models. With good-fitting items as the null hypothesis, a significant difference ($p < 0.05$) means a poor item fit. Particularly, both with and without the Bonferroni correction ($p = \alpha$ / number of items = 0.05 / 12 = 0.0042) were separately applied to detecting misfitting items. RMSEA was simultaneously considered to gauge the magnitude of item misfit, particularly given the sensitivity of the measure when a sample is large (Embretson & Reise, 2000, Orlando & Thissen, 2000). Person fit was calculated based on the *lz* statistic (Drasgow, Levine, & Williams, 1985), denoted by *Zh* hereafter. Examinees with large *Zh* values were considered misfitting, with the absolute *Zh* value of 2.0 or larger as moderate misfit and the absolute value of 3.0 as a severe outlier.

IRT has four key assumptions: unidimensionality, local independence, monotonicity, and item invariance. Unidimensionality means the items on a scale (e.g., 12 EI items) measure one construct in common (e.g., semantic and grammatical accuracy). For the current data, unidimensionality was examined using a parallel analysis for each sample by form and scale type. EFA results were also considered to address the unidimensionality assumption.

The assumption of local independence requires each and every item on a scale to be statistically independent except for the relationships due to the common construct. In other words, residuals of the items should not be (meaningfully) correlated. Although in some contexts, the assumption of local independence can be assumed when unidimensionality is assured in a polytomous unidimensional test (Crocker & Algina, 1986; Ostini et al., 2014), for the current study, the LD-X^2 value for each pair on each scale was examined. Generally, the absolute LD-X^2 value of 10 or greater is flagged for dependence.

To meet the assumption of monotonicity, the probability of the endorsing an item should be on the continuous increase corresponding to individuals' ability levels. For example, examinees with higher L2 proficiency are supposed to score higher on the 12 EI items. This can be examined via person fit, which is also related to item fit.

The final assumption, item invariance, means that estimates of item parameters do not differ regardless of examinee populations, which can be examined via a differential item functioning (DIF). The current project did not directly check this assumption, but the comparison of the entire sample (N = 779) and randomly selected small samples for RQ4 indirectly addressed the issue. The future study is invited to examine DIF for the ACE-EI test.

3.3.3 Random Forest (RF)

RF regression (Breiman, 2001) is the main approach used in Phase II—MLM and univariate regression are used for supplementary purposes, as well. RF regression is a nonparametric statistical analysis based on a series of regression tress using randomly bootstrapped samples, which is referred to as an ensemble procedure. Random forest regression has some advantages over single decision tree-based modeling or traditional linear regression. The single prediction estimated by averaging the predictions of numerous single trees, as well as counter validation of training set against the testing set, increases the predictability and generalizability and avoids overfitting, particularly when a model deals with a large number of predictors compared to the sample size (Grömping, 2009, Hastie, Tibshirani, & Friedman, 2009; Strobl, Malley, & Tutz, 2009). Due to these advantages, RF has appealed to L2 studies that examine a large number of linguistic features in corpus linguistics (e.g., Deshors, 2020, Deshors & Gries, 2020) or language testing (e.g., Fitzgerald et al., 2015). That is, RF is robust to the statistical issues that observational data with a wide range of linguistic features commonly encounter, for example, unbalanced sample size, sparse feature distribution, and numerous, correlated predictors (Gries, 2019). RQ6 of the current study, which aimed to identify the best predictors of item parameters, among numerous linguistic variables, benefitted from RF regression.

RF analysis is conducted with two stages, first with the training data and then with the testing data, usually with the ratio of 2:1. When running a model with the training data, a set of conditions are selected for modeling. Main model specifications include the number of trees to build (i.e., mtree), the number of selected predictors (i.e., mtry), minimum node size of each tree, and the percentage of randomly selected sample. Beginning with a baseline model with default conditions, a series of follow-up RF models are conducted to find the optimal conditions. The optimal conditions from diverse approaches (for example, using different *R* packages to run RF models) are compared to identify the best performing model. The best-performing model is selected based on model performance mostly on the testing set, which is assessed by variance explained (on the training set), predictability (i.e., correlations between predicted values and empirical values), and accuracy vales such as mean square error (MSE) and root mean square error (RMSE) (Breiman, 2001). Smaller values of MSE or RMSE indicated better performance. Appropriate MSE or RMSE values depend on the scale range, mean scores, and/or distribution of the outcome.

One benefit of RF modeling is the variable importance, that is, how important each variable is for prediction of outcome. Variable importance of each predictor is calculated estimating the amount of increase in MSE in trees when a variable of interest is replaced by random noise. For the current study, the increase in MSE was normalized using standard deviation, and thus expressed in Z-score format. Higher values indicate greater importance. If interaction is of interest, the magnitude of the interaction effects can be assessed by the difference between the sum of variable importance of two individual variables and that of a pair. The following sections describe the methodological procedure by research question in each phase.

3.4 Methods by Research Questions

The current study addressed six RQs. Specific research methods for each RQ are summarized in Table 3.6 and elaborated in the following sub-sections.

Table 3.6 Summary of Methods and Dataset Used for Each Research Question

Research questions (Chapter presenting results)	Analysis and dataset
1. What is the optimal number of EI response categories for EI measurement qualities, particularly regarding the independent use of exact repetition and appropriate paraphrase? (Chapter 4)	
1.1 Does a 4-category EI scale, which collapses the lowest category (i.e., no, incomprehensible, or irrelevant response) into its adjacent category (i.e., response with major grammatical errors or meaning difference), more optimally measure accuracy than the current 5-category scale? 1.2 Does an IRT model on the 4-category EI scale support the higher ordering of exact repetition than errorless paraphrase and the use of the two separate categories?	IRT (Graded Response Modeling, GRM) - Agreed / adjudicated scores of all subjects (N = 779) were analyzed - Test, item, person statistics (i.e., fit, parameters) and category characteristics curves were compared
1.3 What item characteristics are associated with the appropriateness of using the two separate categories?	Pearson correlation - Item parameters and adequacy of the paraphrase category (obtained from RQ1.1) were correlated
2. Which examinee ability and response category levels have the largest number of person and item misfits? (Chapter 5)	
2.1 Where did the person misfits of the 4- and 3-category EI accuracy scales most occurred in terms of person ability and response category levels? 2.2 Where did the item misfits of the 4- and 3-category EI accuracy scales most occurred in terms of person ability and response category levels?	Descriptive statistics - Frequencies of unexpected scores (RQ2.2) and unexpected score patterns (RQ2.1) by person ability and response category levels were examined
3. What are the potential sources of person / item misfits and category inadequacy? (Chapter 6)	

Table 3.6 continued

3.1 What are the potential sources of the person misfits on the 4- and 3-category EI accuracy scales in relation to rating grammatical errors and semantic deviations? 3.2 What are the potential sources of the item misfits on the 4- and 3-category EI accuracy scales in relation to rating grammatical errors and semantic deviations? 3.3 Do the issues with rating grammatical errors and semantic deviations observed in misfitting items and examinees exist in item responses with unexpected scores of non-misfitting items or examinees? 3.4 What are the potential source of the inadequacy of using the paraphrase category? 3.5 What guidelines for item development, scale/rubric revision, and rater training do the qualitative analysis provide to minimize misfits and increase the adequacy of using the paraphrase category?	Qualitative analysis of coding sources
4. What is the optimal EI scale / scoring method for measuring L2 semantic and grammatical accuracy? (Chapter 7)	
4.1 What are the alternative EI accuracy scales and rubrics that address the issues of rating criteria in semantic and grammatical judgement?	Qualitative analysis
4.2 Do the quantity-based, alternative EI ordinal scales and rubrics perform better than the quality-based, original EI ordinal scales and rubrics? 4.3 Does the best fitting ordinal scale perform better than binary scale?	IRT (GRM & 2PL) - Using stratified sampling, 360 subjects (90 per from) were selected out of the main subjects (N = 779) - Scores were recoded based on the alternative scales, resulting in 9 sets of scores in total. - Test, item, person statistics (i.e., fit, parameters) were compared among the 9 scales.
5. What are the relationships between EI prompt length and EI item parameters (i.e., item difficulty and item discrimination) across scales / scoring methods? (Chapter 8)	

Table 3.6 continued

5.1 To what extent, does prompt length (i.e., number of syllables, number of words) impact item difficulty across eleven different scales / scoring methods?	Two univariate multiple regressions - DV: item difficulty - IV: number of syllables, number of words - Item difficulty from the scores on the original scale (from RQ1) and the best-performing scale (from RQ4) were analyzed, respectively
5.2 (a) To what extent, does prompt length (i.e., number of syllables) impact item discrimination across ten different ordinal scales / scoring methods? (b) Which characteristics of scale modification are associated with the relationship?	Multi-level modeling (Random intercept model) - DV: item discrimination - IV: number of syllables - Item discrimination from the original scores (from RQ1) and alternative scores (from RQ4) analyzed - scale / scoring method as a cluster
6. Which prompt linguistic features are most important to predicting EI item parameters (i.e., item difficulty and item discrimination) across response category levels and scales / scoring methods? Are there any important interactions? (Chapter 9)	
6.1 (a) Which linguistic features of EI prompts are important to predicting EI item difficulty across the category levels and different scales? (b) Which features interact with category levels of difficulty and/or scales / scoring methods?	Random forest regression - DV: item difficulty (RQ6.1) and item discrimination (RQ6.2) - IV: linguistic features, scales, (and category levels for RQ6.1) - Item difficulty and discrimination from the 9 scales (from RQ4) were analyzed

3.4.1 Adequacy of EI Response Category (RQ1)

RQ 1 was concerned with the optimal number of EI response categories for measurement qualities, with a focus on the independent use of the two highest categories: exact repetition and (appropriate) paraphrase. To address this question, in addition to the current scale (i.e., a 5-category scale that uses exact repetition and (appropriate) paraphrase), two more scales were created: a 4-category EI scale, which collapsed the lowest category (i.e., no, incomprehensible, or irrelevant response) into its adjacent category (i.e., response with major grammatical errors or meaning difference), and a 3-category scale, which combined exact repetition and paraphrase of the 4-category scale. Twelve univariate IRT models in total were constructed and run onto the item scores obtained from each form of the three scales.

In order to check the extent to which the lowest category is appropriate to the measurement purposes, measurement qualities were compared between the current 5-category scale and the 4-category scale (RQ1.1). Measurement qualities examined were test- and item-level indices and fit statistics mentioned in Section (3.3.2).

Next, the IRT models based on the 4-category scale were examined to see if the item difficulty of exact repetition was higher than item difficulty of paraphrase across the 48 items using the ordering of the curves in the CCCs and threshold values (RQ1.2a). If yes, the IRT-based empirical results support the conceptual approach of higher proficiency demonstrated by exact repetition over paraphrase. Also, the category curves in the CCCs of the 48 items were examined to see if each category of each item had a clear peak with a sufficient range without overlapping with the range of another category's curve (RQ1.2b). Categories without a clear peak or sufficient independent range means lack of the adequate usage of the category.

Finally, item characteristics associated with the appropriateness of using the two highest categories were identified via correlational analysis (RQ1.3). The appropriateness was indicated by two variables: 1) the total probability of the endorsement by the paraphrase category (i.e., the sum of values of the fitted curve on a CCC for a given latent trait range, theta -4.0 to 4.0) and 2) illustrated practical value of the paraphrase category (i.e., 2: a clear peak with sufficient non-overlapping range, 1: a peak with a small non-overlapping range, 0: little or no non-overlapping range). The correlated item characteristics were item discrimination (a), threshold of paraphrase (b_2), threshold of exact repetition (b_3), overall difficulty ($b_{overall}$), distance between b_2 and b_3, and proportion of the number of categorical responses (frequency of responses in the item category / total number of responses in the item × 100).

3.4.2 Misfits Across Examinee Ability and Response Category (RQ2)

Among the measurement qualities, RQ2 focused on misfit distribution by examinee ability and response category levels for three misfitting items (RQ2.1) and 22 examinees (RQ2.2) respectively on the 4- or/and 3-category EI scales. Frequency analysis was used to examine the relationships based on the differences between IRT-model based scores and empirical scores, in addition to the item and person fit statistics. Also, interaction plots of and item performance plots were used to graphically examine the relationships.

3.4.3 Potential Sources of Misfits and Category Inadequacy (RQ3)

Expanding on the quantitative misfit analysis from RQ2, RQ3 qualitatively examined the actual examinee responses with unexpected scores (i.e., scores that show a difference of | ± 1.0| or larger than model-based expected item scores) in search of the potential sources of person misfits (RQ3.1) and item misfits (RQ3.2). Based on the researcher's experience of rater training and rating, as well as rater justifications, the qualitative coding was focused on the patterns of inconsistency of applying grammatical and semantic rating criteria in relation to the types of omission and paraphrasing. Also, responses with unexpected scores of non-flagged items and examinees were examined to see if the patterns detected from the flagged items and examinees can be broadly applied (RQ3.3). With a similar approach, responses with unexpected item scores in the paraphrase category were examined to find the potential source of the inadequacy of using the paraphrase category (RQ3.4). Based on the results, suggestions for item development, scale/rubric revision, and rater training were provided to minimize misfits and increase the adequacy of using the paraphrase category.

3.4.4 Revised Scales and Scoring Methods (RQ4)

Using the advisory information on the sources of misfit from RQ3, RQ4 created alternative scales / scoring methods (RQ4.1) and IRT model performance of the alternative ordinal scales were examined (RQ4.2). Three alterative scoring methods were proposed to revise the rating criteria for the current minor and major deviation categories. The modified rating criteria were 1) frequency-based semantic deviation (FSD, one to four semantic deviations versus more than four) instead of minor versus major semantic deviation, 2) frequency-based grammatical deviation (FGD, one or two grammatical errors versus more than two errors) to further categorize the minor grammatical error category, and 3) the combination of both frequency-based semantic and grammatical deviation (FSGD). These three methods were applied to 3-category and 4-category ordinal rating scales, which resulted in six alternative ordinal scales / scoring methods. A binary scale was created by collapsing the major and minor deviation categories of the 3-category scale. The seven alternatives and two 3-category and 4-category scales without modification except for collapsed categories, the total of nine options, were used to explore the most optimal EI scales / scoring methods to measure semantic and grammatical accuracy.

Ninety examinees per form, the total of 360 examinees were selected from the entire sample (N = 779), based on stratified sampling. The 360 examinees' original item scores were rescored based on each of the seven alternatives. The rescoring resulted in 9 sets of scores per examinee. Details of the revised criteria of the alternative scales and scoring methods, including examples, are found in Chapter 7 (see Section 7.1). A univariate GRM model was run for each ordinal option and 2PL model for the binary alternative. IRT model performance, including measurement and model / item / person fit statistics (See Section 3.3.2) were compared to identify the best performing, and thus most optimal scale / scoring method.

3.4.5 Association Between EI Prompt Length and EI Item Parameters (RQ5)

RQ5 and RQ6 examined the relationships between prompt features and EI item parameters (Phase II). Particularly, RQ5 investigated the impact of the prompt length (i.e., number of syllables, number of words) on EI item difficulty (RQ5.1) and item discrimination (RQ5.2) across the scales and scoring methods. For the analysis of RQ5.1, only the item difficulty at the paraphrase level was examined, which was of the greatest interest because the threshold is potentially most crucial for the program's course exemption decision and assessment of L2 proficiency development over the courses. Three univariate regression models were run onto three different samples, with each having the prompt length (i.e., number of syllables) as a predictor and item difficulty as an outcome variable. The three outcome groups used for the univariate analysis were: 1) item difficulty of 3-category and 4-cateogry scales without modification based on the entire sample (N = 48 × 2 = 196), 2) item difficulty of all scales, including the two sets from the entire sample (N = 48 × 11 = 528), and 3) item difficulty of the 4-category FSGD scale (N = 48). These samples were selected because the first sample reflects the performance of the non-modified scoring methods from the large groups, the second is most comprehensive, and the third is the best-performing model from RQ 4. Prediction of another measure of prompt length, the number of words, was compared by running three additional univariate regression models with the same set of data. Note that the item difficulty values were nested within the scales / scoring methods, but MLM was not considered because the proportion of variation explained by the scales / scoring methods (a.k.a., ICC) was marginal.

For the analysis of item discrimination, values from all scales / scoring methods, including 3-category and 4-cateogry scales without modification applied to the entire sample, were collected, except for the binary scale because of the precision issues with item discrimination of the binary scale detected in RQ4. Due to the nesting data structure, MLM was used. Unlike the analysis of item difficulty, the proportion of the variance in item discrimination explained by scales / scoring methods was not negligible from the ICC value of the null model. Thus, two-level MLM was used to examine the association of prompt length and item discrimination with item discrimination across scales / scoring methods. Beginning with a random intercept model as a baseline model, where item discrimination and prompt length (i.e., the number of syllables) are Level-1 outcome and predictor respectively, three Level-2 predictors (i.e., frequency-based semantic rating, frequency-based grammar rating, and collapsing paraphrase into exact repetition) were added one by one to identify important characteristics of scale modification. For the importance characteristics, the cross-level interaction was examined. Predictors (i.e., fixed effects) were assessed with a p-value ($p < 0.05$), and model comparison were made based on proportion reduction between and within scale residuals, chi-square difference test, and changes in BIC and AIC values.

3.4.6 EI Prompt Linguistic Features Important to Predicting EI Item Parameters (RQ6)

RQ6 examined prompt linguistic features in search for most important contributors to predicting EI item parameters, item difficulty (RQ6.1) and item discrimination (RQ6.2) across response category levels and scales / scoring methods, including important interactions. First the syntactic and lexical linguistic features of EI prompt were extracted by using NLP-based open-source programs: the Tool for the Analysis of Lexical Sophistication (TAALES 2.0) (Kyle et al., 2018) for lexical sophistication indices, and the Tool for the Automatic Assessment of Syntactic Sophistication and Complexity (TAASSC, Kyle, 2016). Morphological features were manually coded. Extracted values, including prompt length (i.e., number of syllables), were examined for correlations. If a pair of valuables is highly correlated ($r \geq 0.7$), the variable with lower representativeness, fewer frequencies, or/and lower correlation with the outcome variable (i.e., either item difficulty or item discrimination) was deleted. Also, only the indices with correlation of 0.3 or higher with the outcome variable) were included. This process left 47 linguistic variables, which are listed in Appendix 3.1. Two non-linguistic variables were included, as well: (eleven) scales / scoring methods (i.e., seven alternatives, 3-category and 4-category scale without modification with large and small samples, respectively) and (five) item category levels (i.e., threshold levels: exact repetition, paraphrase, (upper, all) minor deviation, major deviation). Note that the item discrimination from the binary scale was excluded for modeling because of the precision issue.

Item difficulty and item discrimination were outcome variables, which were examined independently. Each of the item parameters was predicted at three different levels respectively: at the paraphrase, minor error / deviation, and all levels. The models that predicted item difficulty or item discrimination at the high (i.e., the paraphrase level) and low level (i.e., minor error / deviation level) were compared with each other to examine the level-specific performance of linguistic variables for the prediction. The models at all levels disclosed overall patterns.

Following the procedure described in Section 3.3.3, a series of RF regressions were conducted, and six best performing RF model was identified for the prediction of each item parameter at the three threshold levels (i.e., paraphrase, minor errors / deviation, and all levels), respectively. The *randomeForest* R package (Liaw & Wiener, 2002) and the *ranger* R package (Wright & Ziegler, 2015) were used for RF models. Important linguistic prompt features (and non-linguistic features) were identified using the values and ranks of variable importance of each predictor from the six best-performing models. For another set of six optimal models with interactions, the randomForestSRC R package (Ishwaran, & Kogalur, 2014) was employed. From the best-performing models, the magnitude of the interaction effects of all possible pairs were calculated, which is the difference between additive and paired variable importance. Large difference was considered potentially important interactions. Also, plots of marginal effects were compared between the three models of different levels. Different patterns were considered indication of potential interaction between the given linguistic feature and threshold levels.

PHASE I. MEASURING SEMANTIC AND GRAMMATICAL ACCURACY: OPTIMAL SCALES/ SCORING METHODS AND MISFIT ANALYSIS

The study investigated ACE-In EI as a measure of L2 English proficiency in two phases: exploring optimal EI scales/ scoring methods, including the sources of misfits (Phase I) and 2) examining the impacts of prompt features on EI measurement qualities (Phase II). In Phase I, four main topics were examined in each of the four chapters regarding the optimal EI scales, scoring methods, and rubrics: 1) the adequacy of EI response categories, 2) person and item misfits across examinee proficiency and response category levels, 3) potential sources of misfits and category adequacy, and 4) revision of scales and rubrics. Note that the EI test has four forms, with each consisting of twelve items. Thus, each form was analyzed separately when IRT models were applied.

CHAPTER 4. ADEQUACY OF EI CATEGORIES (RQ1)

As the first step of exploring an EI scale to optimally measure overall accuracy, the current ACE-In EI scale was examined, focusing on the adequacy of the scale categories (RQ1). In this section, I presented the results of a preliminary analysis of CTT-based descriptive statistics (Section 4.1) and the main IRT analyses that fit GRMs to scores on three accuracy scales: 5-, 4-, and 3-category scales (Section 4.2 to 4.5), including preliminary tests to establish IRT assumptions, that is, unidimensionality and local independence. Specifically, the following questions were answered.

RQ1.1 Does a 4-category EI scale, which collapses the lowest category (i.e., no, incomprehensible, or irrelevant response) into its adjacent category (i.e., response with major grammatical errors or meaning difference), more optimally measure accuracy than the current 5-category scale? (Section 4.2)

RQ1.2 Does an IRT model on the 4-category EI scale support the higher ordering of exact repetition than errorless paraphrase and the use of the two separate categories? (Section 4.3)

RQ1.3 What item characteristics are associated with the appropriateness of using the two separate categories? (Section 4.4)

RQ1.4 Does a 3-category EI scale, which collapses errorless paraphrase into exact repetition, more optimally measure accuracy than the 4-category scale? (Section 4.5)

The final section (Section 4.6) discusses implications for EI item development, the use of EI scores, and related research.

4.1 Descriptive Statistics of EI Accuracy Scores

Appendix 4.1 and Table 4.1 present descriptive statistics of the EI scores from the original 5-category ordinal scale with the range of 0 to 4 that measured grammatical and semantic accuracy. Although IRT modeling for this study assumes ordinal responses and the current scale was ordinal, the descriptive statistics included item means, standard deviations, and normality indices for a general understanding of the data, following the convention that allows for treating scores measured on an ordinal scale with five or more categories as an ordinal approximation of a continuous variable response (Norman, 2010; Sullivan & Artino, 2013; Zumbo & Zimmerman, 1993). For the majority of the items, 38 out of 48 items, the item median was 2 while four items showed a median of 3 (i.e., Item 3, 18, 31, and 32) and six items had a median of 1 (i.e., Item 10, 11, 13, 25, 26, and 45). Item means across the four forms ranged from 1.10 (Item 10) to 2.95 (Item 18), showing considerable variation among items. All items were within the acceptable normal range of skewness or kurtosis values, which were below the absolute value of 1.5 (Tabachnick & Fidell, 2013), except for two items: Item 11 (kurtosis 1.88) and Item 30 (kurtosis -1.68). In Table 4.1, the ranges of item skewness and kurtosis seemed comparable across the four forms, except for the kurtosis of Form 1, which has largest item kurtosis (i.e., Item 11). Standard errors (SE) were small across the forms, being less than 0.1. Average item and total scores also appeared quite similar across the forms, with Form 3 being slightly higher, but their ranges varied to some degree.

Regarding the range of the scale used, the full five-category range was used in only two thirds of the items, and 16 items had no responses to the lowest score (see the Range column in Appendix 4.1), which indicates the potential need for collapsing categories. Table 4.2 shows the number of items of which categories had a small percent of the total responses, 5% and 10% or less. Appendix 4.2 presents a detailed description—the number and proportion of responses for each category of the 48 items. Including 16 items that had no response rated as zero, 45 out of 48 items had responses of 5% or less in the category of score zero. The other extreme end, the category of the highest item score (i.e., 4; exact repetition) was second least used. Slightly more than one third of the items had responses of 10% or less, including 13 items with responses of 5% or less. In the category of item score 3 (i.e., errorless paraphrase), approximately one fourth of the items had 10% of total responses or less, and six items, 5% or less. Only 21 items, less than the half of the total number of items, had responses more than 10% in both the exact repetition and errorless paraphrase categories.

Table 4.1 Descriptive Statistics of the EI Item Scores by Forms (the 5-Category Accuracy Scale)

Form	Mean		Range (min., max.)				
	Item Score	Total Score	Item Mean	Total Score	Item Skewness	Item Kurtosis	Item SE
Form 1	2.10	25.24	1.68	31	0.92	3.34	0.03
(N=193)			(1.10, 2.79)	(11, 42)	(-0.18, 0.74)	(-1.46, 1.88)	(0.05, 0.08)
Form 2	2.12	25.87	1.61	39	1.19	2.63	0.03
(N=202)			(1.34, 2.95)	(8, 47)	(-0.34, 0.85)	(-1.45, 1.18)	(0.05, 0.08)
Form 3	2.19	26.32	1.27	37	1.03	2.42	0.04
(N=204)			(1.62, 2.89)	(8, 45)	(-0.18, 0.85)	(-1.68, 0.74)	(0.05, 0.09)
Form 4	2.16	25.91	1.24	30	0.77	2.30	0.04
(N=180)			(1.53, 2.77)	(13, 43)	(0.02, 0.79)	(-1.42, 0.88)	(0.05, 0.09)
Total	2.14	25.71	1.84	39	1.19	3.56	0.04
(N=779)			(1.10, 2.95)	(8, 47)	(-0.34, 0.85)	(-1.68, 1.88)	(0.05, 0.09)

Note: Form 1: Item 1 to 12; Form 2: Item 13 to 24; Form 3: Item 25 to 36; Form 4: Item 37 to 48

Table 4.2 The Number of Items for Low Frequency Categories

Form	**5% or less**					**10% or less**				
	0	**1**	**2**	**3**	**4**	**0**	**1**	**2**	**3**	**4**
Form 1	11	0	0	2	4	11	0	0	3	4
Form 2	11	0	0	1	4	11	1	0	3	5
Form 3	11	0	0	2	2	12	1	0	2	3
Form 4	12	0	0	1	3	12	2	0	2	6
Total	45	0	0	6	13	46	4	0	10	18

Note: Form 1: Item 1 to 12; Form 2: Item 13 to 24; Form 3: Item 25 to 36; Form 4: Item 37 to 48

Low frequency of the extreme categories suggested the potential need for collapsing the categories into their adjacent categories, particularly for the lowest category, which was rarely endorsed (i.e., selected) across the items. The tendency in the exact repetition category, however, was not as consistent throughout the items as in the lowest item score category, which complicates the discussion. The adequacy of collapsing the two extreme categories will be discussed in the next two sections, focusing on the comparison between the 5- and 4-category scales (Section 3.3.2) and the 4- and 3-category scales (Section 3.3.3)

4.2 Adequacy of Using the Lowest Category on the EI Accuracy Scale: Comparison Between 5-Category and 4-Category Scales (RQ1.1)

In the current sample, the majority of the items had no or too few responses in the lowest category (i.e., no/incomprehensible/irrelevant response) with only two exceptions, Item 10 and Item 13. For measurement precision (Embretson & Reise, 2000; Ostini et al., 2014), the lowest category was collapsed into its adjacent category (i.e., response with major grammatical errors or major semantic deviation). To evaluate the adequacy, I examined whether or not the 4-category EI scale, which combined the two lowest categories, more optimally measure accuracy than the current 5-category scale (RQ. 1.1a). I particularly attended to Item 10 and 13, because the two items had more than insufficient number of responses in the lowest category, 17.10% and 14.85% of the total, respectively. Results from GRM modeling supported the 4-category scale as a good/better alternative of the 5-category scale by demonstrating comparable or better statistics at test, item, and person levels, particularly regarding SE values of item difficulty parameters and item fits. Details were presented below, including results from testing IRT assumptions.

IRT Assumptions

Prior to fitting GRM, the IRT assumptions of unidimensionality and local independence were checked for both 5- category and 4- category (grammatical and semantic) accuracy scales. The unidimensionality assumption was particularly important for these scales because the scores combined grammatical and semantic accuracy.

The parallel analyses indicated that all four forms on both 5- and 4- category scales met the assumption of unidimensionality (See Appendix 4.3). Although meeting the assumption of unidimensionality is generally assumed equivalent to satisfying the assumption of local independence in a polytomous unidimensional test (Crocker & Algina, 1986; Ostini et al., 2014), examination of the LD-X^2 value for each pair on both scales revealed slight local deviations (See Appendix 4.4). On the 5-category scale, eight out of the total 264 pairs showed the absolute LD-X^2 value of 10 or larger, two of which were larger than 20. With the 4-category scale, only one pair (i.e., Item 13 and Item 22) was flagged, and the value (12.01) was only slightly larger than the cut-off. The large values of the flagged pairs of the 5-category scale are highly likely due to the data sparseness in the lowest category, rather than actual local dependence (Cai, du Toit, Thissen, 2011). With only one locally-dependent pair, scores on the 4-category scale indicates higher statistical stability than on the 5-category scale for the current data.

4.2.1 Test Statistics

The GRM models fitted to scores on the 5- and 4-category scales were compared based on the four qualities of measurement statistics: overall model fits, overall factor loading, test information, and test reliability.

Model Fit and Overall Factor Loadings

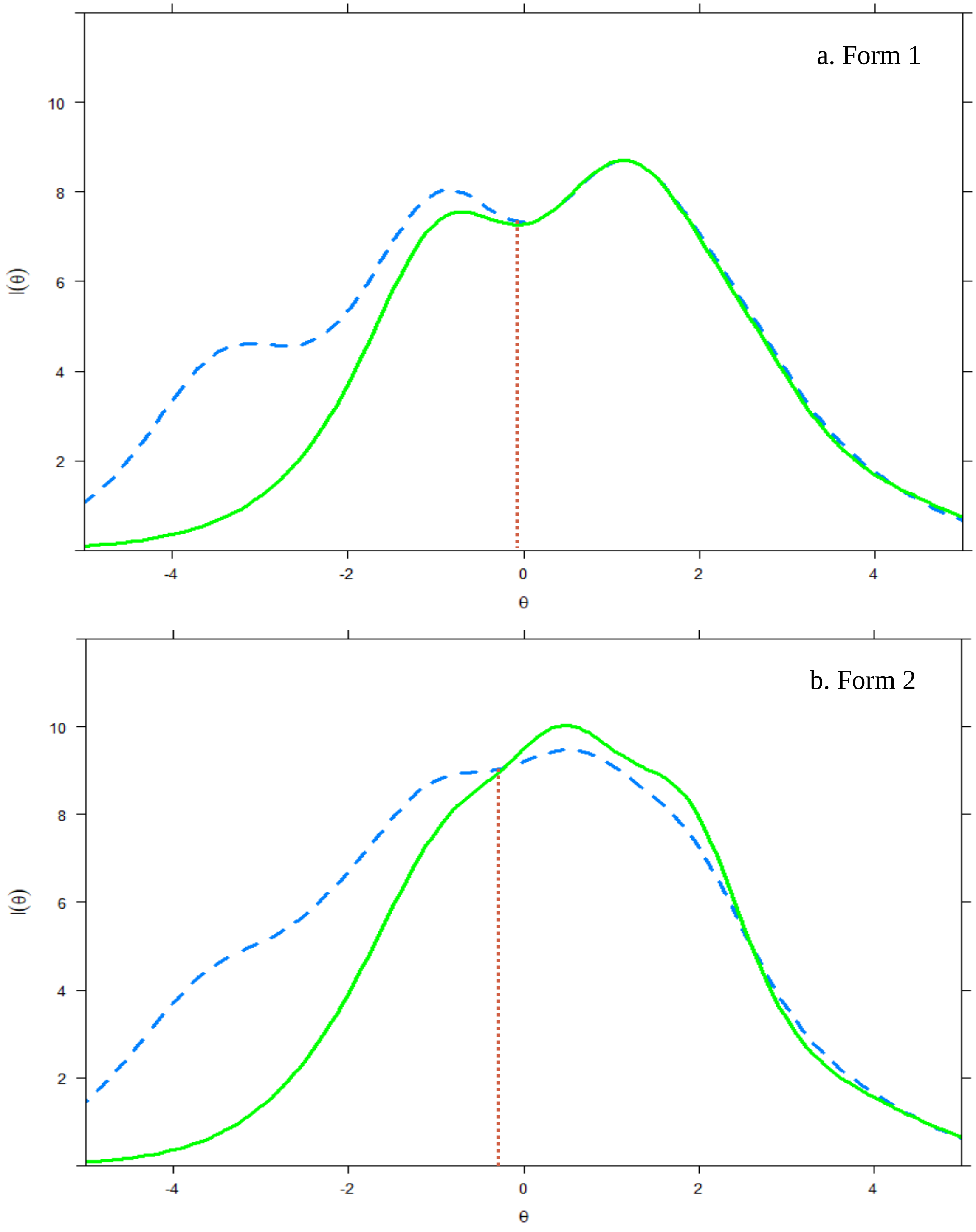
The overall model fits of the GRM models on the 5- category and 4- category accuracy scales were examined based on M_2 statistics. The models on both scales generally fit well. Except for Form 4 (i.e., Item 37 to Item 48), all the *p*-values were insignificant, and TLI and CFI values were 0.96 or higher, and RMSEA value 0.02 or lower, all of which demonstrate good fit (See Appendix 4.5). Form 4 on the 5-category scale did not fit well (*p*=0.01; RMSEA=0.07; TLI=0.85; CFI=0.90) but the fit was improved on the 4-category scale to an acceptable (*p*=0.05; RMSEA=0.07) or a good fit (TLI=0.96; CFI=0.97). When comparing information criteria, the reduced model decreased AIC and BIC values, which is expected because of the fewer number of parameters on the 4-category (i.e., 48) than on the 5-categroy scale (i.e., 55 to 57). Similarly, the sums of factor loadings, which represent the sum of relationships of items and the latent variable, and proportion of variance explained were, although marginally, increased on the 4-category scale across all four forms (See Appendix 4.6). In sum, collapsing the lowest category into the adjacent category yielded as good or more favorable overall model fit, representing the data slightly better.

Test Information and Conditional Reliability

Figure 4.1 and Figure 4.2 illustrate the test information and reliability of EI accuracy across different theta (i.e., ability) levels of the two scales. The dotted and solid curves respectively represent the estimates on the 5- and 4-category scales. The dotted straight lines point out the value where the two curves met. Generally, scales with a larger number of categories are expected to provide more information (de Ayala, 2013). Because the two lowest categories of the 5-category scale were combined into one category on the 4-category scale, the information that the instrument provided was reduced at the low theta levels of the 4-category scale across the four forms.

Approximately at the minimum theta level -1.2 (Form 4) and lower—and -0.2 (Form 1) and lower at maximum—the 5-category scale provided more information than the 4-category scale. At the rest of the theta levels, the two scales yielded almost the same amount of information. Interestingly, Form 2 displayed slightly larger information at the theta level -0.3 or higher on the 5-category than the 4-category scale. Overall, the amount of reduced scale information might appear large, but considering the number of examinees at those low theta levels is marginal, the practical difference would be very little. Apart from the main research questions, however, the information for Form 4 was found to be smaller than the other three forms, which will be discussed later.

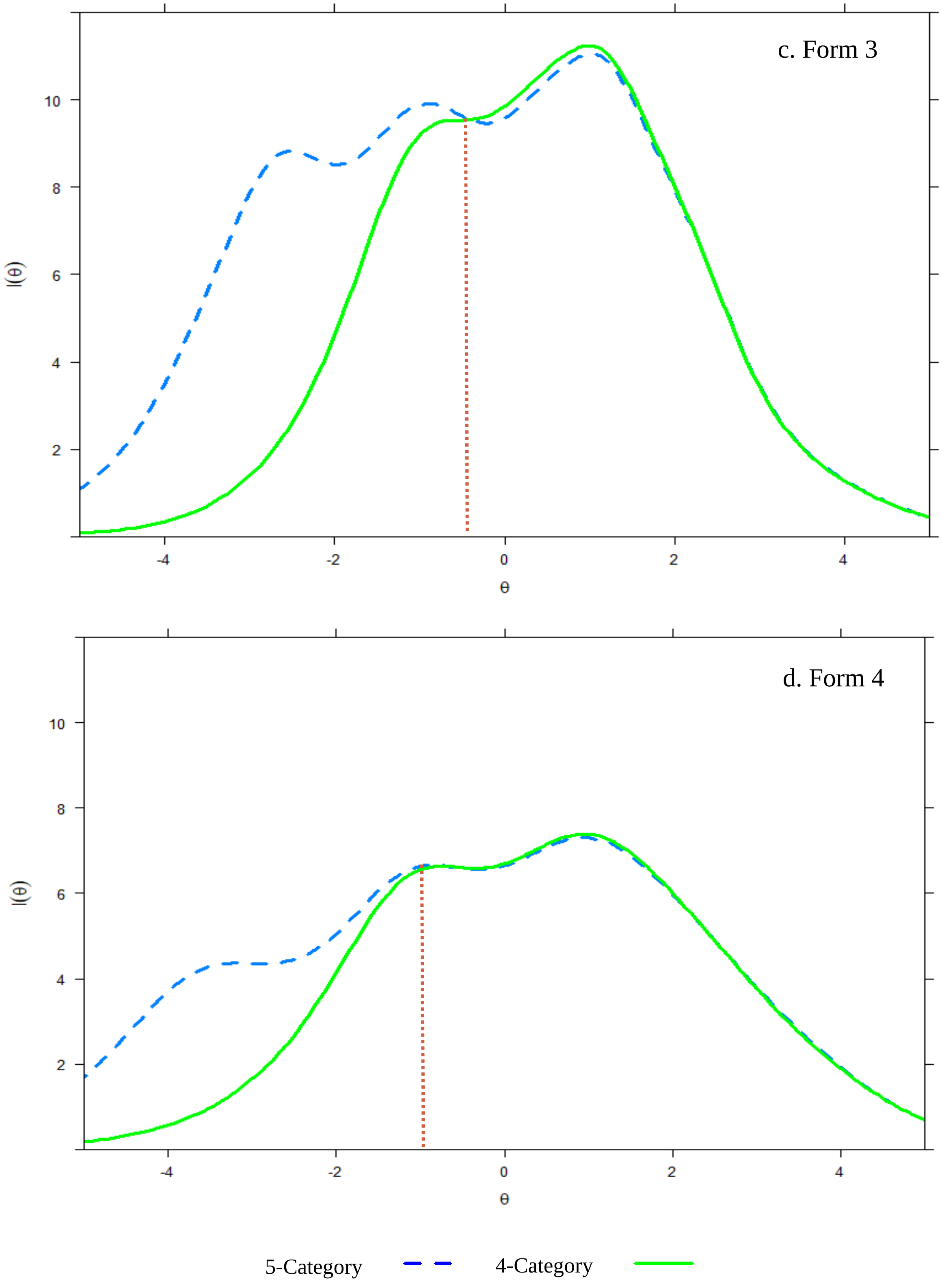
Similarly, the conditional reliability of the 5-category scale is higher than of the 4-category scale at the theta level -1.6 (Form 4) or lower at the minimum and -0.8 (Form 2) or lower at the maximum. It is not surprising to observe that the distribution of the conditional reliability corresponds to that of the scale information because conditional reliability is mathematical transformation of function as information and standard errors across different theta levels. Since the target population of the scale is not the lower end and the number of examinees of the lowest end is marginal, the decreased reliability at the lower theta levels on the 4-category should not be an issue. Also, it is possible that some target population members may be at the low end of the scale, but the passing total score likely is not at the low end of the total score scale. Moreover, the single IRT reliability estimates of the EI accuracy on both scales were all high with difference less than 0.01 between the scales, ranging from 0.87 (Form 4) on both scales to 0.91 (Form 3) on the 5-category scale. Thus, in terms of (conditional) reliability, the 4-category scale seems a good alternative.


a. Form 1

b. Form 2

Figure 4.1 Test Information Functions for the GRM Models of 5--Category and 4-Category EI Accuracy Scales

Figure 4.1 continued



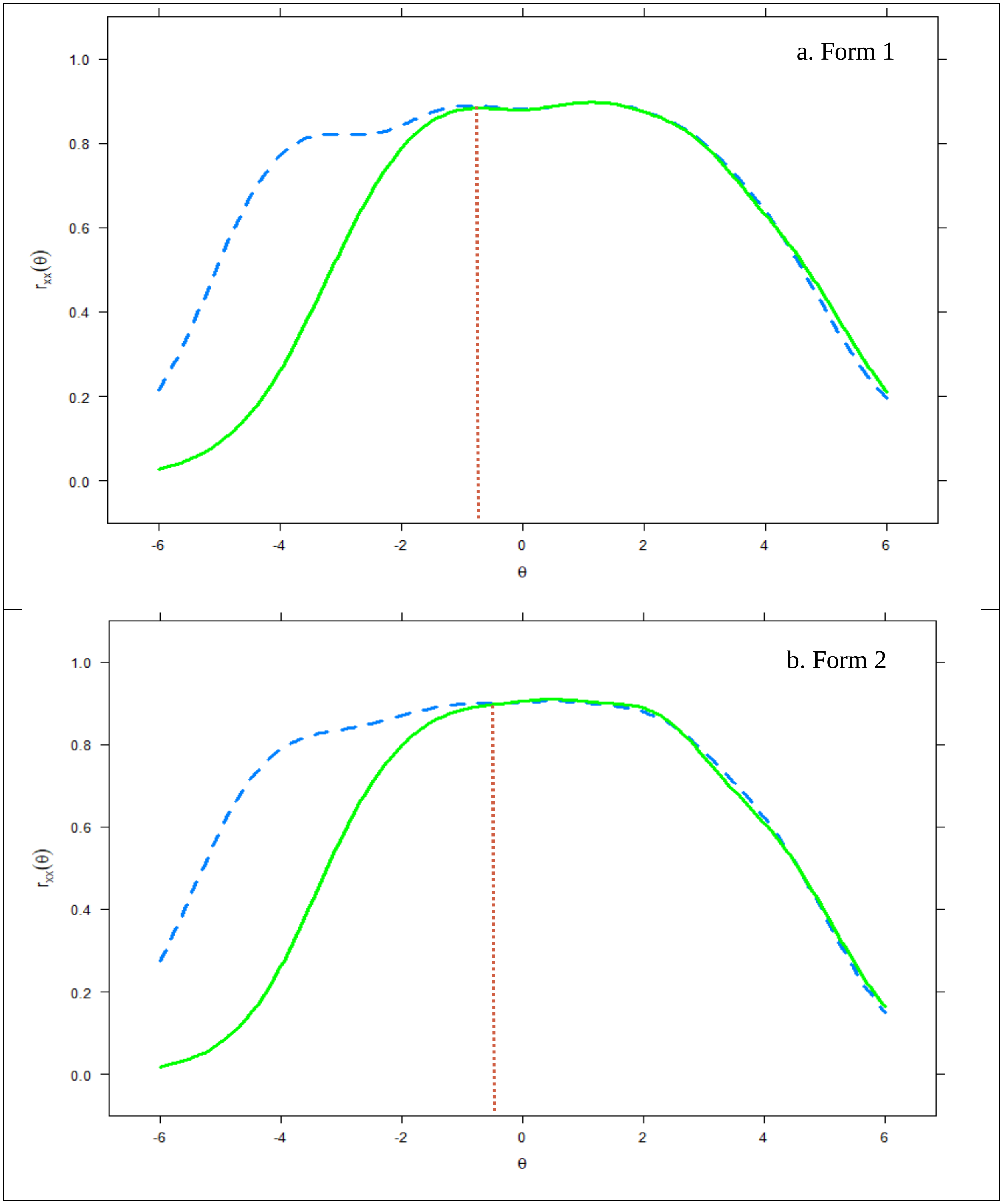
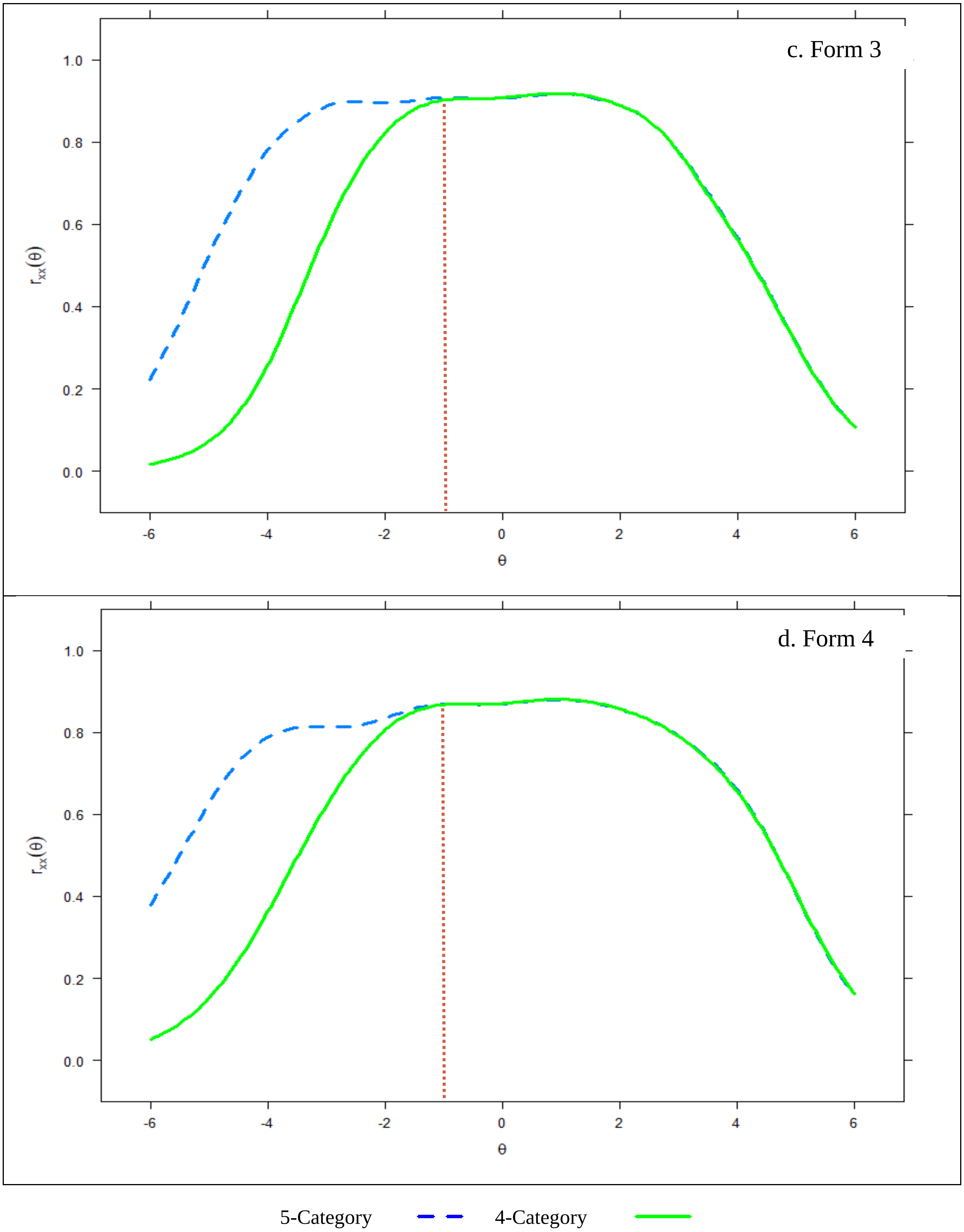


Figure 4.2 Conditional Reliability of the 5--Category and 4-Category EI Accuracy Scales

Figure 4.2 continued



4.2.2 Item Statistics

In addition to test-level statistics, item- and person-level statistics of the GRM models fitted to scores on the 5- and 4-category scales were compared. Three item statistics were compared: item discrimination, item difficulty, and item fit.

Item Discrimination

Table 4.3 shows the distribution of the item discrimination values on both scales based on Baker and Kim's (2017) classification (See Appendix 4.7). Overall, the item discrimination values of both scales ranged from moderate to very high. The minimum and maximum item discrimination of the 4-category scale (Item 11, a=1.02; Item 13, a=2.64) were similar to those of the 5-category scale (Item 11, a=0.94; Item 34, a=2.66). The discrimination estimates on both scales were all acceptably precise (SE <0.4) with only one exception—SE of Item 13 on the 4-category scale was borderline (SE = 0.41).

Table 4.3 The Distribution of the Item Discrimination of the Scores Measured on the 5-Catgory and 4-Category EI Accuracy Scales

Discrimination (a)	No. of Items	
	5-category scale	4-category scale
Very low (a < 0.35)	0	0
Low (0.35 ≤ a < 0.65)	0	0
Moderate (0.65 ≤ a < 1.35)	5 (min. 0.94 (Q11))	8 (min. 1.02 (Q11))
High (1.35 ≤ a < 1.70)	24	24
Very High (a ≥ 1.70)	19 (max. 2.66 (Q34))	16 (max. 2.64 (Q13))
Total	48	48

A comparison of item discrimination between the two scales elaborated in Table 4.4 also showed little differences in general. Most items, 43 out of 48 items made changes of |±0.1| or smaller in their discrimination. There were only five items (i.e., Item 10, Item 13, Item 24, Item 25, and Item 26) showed changes larger than 0.1. It is highly likely that the five items showed clearer impact of combining the two lowest categories because of their (relatively) larger number of responses in the lowest category—the items had the top five largest number of responses (See Appendix 4.2). The larger number of responses allowed for more power to change the discrimination estimates. Similarly, the greater number of responses in the collapsed category, the greater change in the SE of the discrimination estimate. The SEs slightly increased in the five items (i.e., Item 10, Item 13, Item 24, Item 25, and Item 26) with the maximum of 0.04 in Item 10 (from 0.26 to 0.30) and Item 24 (from 0.28 to 0.32) while all the other items showed little change.

Table 4.4 Changes in Item Discrimination (a) Estimates of the EI Scores Measured on 5-Category to 4-Category Accuracy Scales

Change in discrimination ($\Delta(a)$, $a_{5\text{-category}} - a_{4\text{-category}}$)	Frequency	Item
$\Delta(a) > 0.2$	1	Q13[a]
$0.1 < \Delta(a) \leq 0.2$	3	Q24, Q25, Q26
$-0.1 \leq \Delta(a) \leq 0.1$	43	The rest
$-0.2 \leq \Delta(a) < -0.1$	1	Q10[b]
$\Delta(a) < -0.2$	0	N/A
Total	48	

Note. [a] $\Delta(a) = 0.49$; [b] $\Delta(a) = -0.17$

Interestingly, despite the similar descriptive statistics, discrimination of Item 10 decreased from 1.77 to 1.60 ($\Delta(a) = -0.17$) but the parameter of Item 13 notably increased from 2.15 to 2.64 ($\Delta(a) = 0.49$). Item 10 and Item 13 are the items with the top two largest number of endorsements in the lowest category, 33 (17.10%) and 30 (14.85%), respectively, while being the two most difficult items from their lowest means, 1.10 and 1.34. However, Item 10 has more responses in the category 2 (n = 117, 60.62%) and fewer in the category 4 (n = 1, 0.52%) than Item 13 in the category 2 (n = 94, 46.53%) and category 4 (n = 8, 3.96%). These differences led to much higher difficulty estimation in Item 10 ($b_{overall} = 1.48$) than in Item 13 ($b_{overall} = 0.84$) when using IRT modeling. By combining the two lowest categories of Item 10, the number of responses to the collapsed category (n = 150 = 33 + 117) came to account for 77.72% (17.10% + 60.62%), which lowered the item discrimination. On the other hand, the collapsed category of Item 13 (n = 124 = 30 + 94) takes up 61.38% (14.85% + 46.53%), still having fairly large proportion endorsed by the other categories.

Similarly, the EFA factor loadings of the items between the two scales rarely varied or to marginal extent, aligning with the slightly increased sums of factor loadings on the reduced scale (See Appendix 4.6 for details). The items on both scales loaded moderately to strongly, ranging from 0.48 (Item 11) to 0.84 (Item 34) on the current scale, and from 0.51 (Item 11) to 0.84 (Item

13) on the 4-category scale. Similar to item discrimination, the factor loading most increased in Item 13, from 0.79 to 0.84, while most decreased in Item 10 from 0.72 to 0.68.

In summary, the more parsimonious scale did not degrade item discrimination, and precision remained acceptable, which supports the use of the 4-category over the 5-category scale. Item discrimination was mostly consistent between the two scales or slightly increased in some items on the 4-category scale. One exception (i.e., Item 10) existed but the decrease in discrimination and impact for the model was marginal.

Item Difficulty

Unlike discrimination estimates, overall item difficulty values varied between the 5- and 4-category scales, according to Table 4.5 (See Appendix 4.7 for details). Note that, however, boundary thresholds (i.e., category-level item difficulty) rarely shifted. Table 4.5 shows that overall item difficulty generally went up after combining the two lowest categories. The increases occurred in the items that had non-zero responses in the lowest category, with the maximum difference of 1.66 (Item 37). Moving to the 4-category scale reduced the number of the easy items, from 15 to 4, which rendered approximately half of the items moderately hard to very hard.

Table 4.5 The Distribution of Overall EI Item Difficulty Estimates ($b_{overall}$) Measured on the 5-Category and 4-Category Accuracy Scales

Difficulty	No. of Items	
	5-category scale	4-category scale
Very easy ($b_{overall} < -2.0$)	0	0
Easy ($-2.0 \leq b_{overall} < -0.5$)	15 (min. -1.05 (Q19))	4 (min. -0.75 (Q18))
Medium ($-0.5 \leq b_{overall} < 0.5$)	25	22
Moderately hard ($0.5 \leq b_{overall} < 1.0$)	4	8
Hard ($1.0 \leq b_{overall} < 2.0$)	4 (max. 1.48 (Q10)	11
Very Hard ($b_{overall} \geq 2.0$)	0	3 (max. 2.70 (Q11)
Total	48	48

The increase was expected because overall difficulty is the average of boundary locations. When the lowest boundary thresholds were deleted, the average went up. Thus, whether or not the increase is more precise measurement depends on the measurement quality of each boundary threshold, particularly the estimates of the collapsed category (i.e., SE_{b1}) on the 4-category scale.

Importantly, difficulty estimates of the lowest category on the 5-category scale were not always precise. Among the 32 items with non-zero responses in the lowest category on the 5-category scale –16 items had no responses in the lowest category, meaning no boundary threshold estimate for item score 0—23 items had SE larger than 0.4, which is beyond the acceptable precision. The issue was solved when collapsing the two lowest categories. The lowest boundary thresholds on the 4-category scale, or the second lowest on the 5-category scale, were all within the acceptable range (SE < 0.4) except for Item 48 ($SE_{b2,\ 5\text{-category}} = SE_{b1,\ 4\text{-category}} = 0.48$) (See Appendix 4.7 for details).

In contrast to the overall item difficulty, the boundary thresholds were consistent between the two scales with differences less than 0.1 for most items. There were only four items (i.e., Item 10, Item 11, Item 13, and Item 23) for which the difference was greater than the absolute value of 0.1. Table 4.6, however, indicates that the larger differences ($\Delta(b) > |\pm0.1|$) were not made in the lowest thresholds but the rest. Also, the changes made in Item 11 and the highest thresholds of Item 10 and Item 23 were not precise (SE > 0.4). When precision of the thresholds considered, only Item 11 and Item 13, which had the sufficient responses of the lowest category on the 5-category scale, made changes in their categorical thresholds, either slightly decreased (Item 13) or increased (Item 10).

Interestingly, the four items (i.e., Item 10, Item 11, Item 13, and Item 23) had some commonalities, compared to the other 44 items on the 5-category scale. Table 4.7 demonstrates the items had 1) a (relatively) higher frequency of the two lowest categories, respectively and combined, 2) fewer responses in the highest category, 3) lower CTT-based item means, and 4) higher IRT overall difficulty values. The four items also went through a relatively larger change in their discrimination values from the 5-category to 4-category scale. From the common statistical features, the relatively larger differences observed in the difficulty estimates of the four items were likely due to the unstable/inflated difficulty values and larger power of the lower categories from the relatively sufficient number of responses.

In sum, the difficulty of the lowest category on the 5-category scale, the boundary between score 0 and score 1 or above, was not acceptably precise or did not exist in the majority of the items (i.e., 39 out of 48 items). When collapsing the lowest categories into their adjacent categories, corresponding threshold boundaries hardly varied between the 5- and 4-category scales, except for four items, only two of which were precise. Since the item difficulty estimates on the 4-category scale were generally comparable to the estimates on the 5-category scale while excluding thresholds beyond the acceptable range of preciseness, the more parsimonious scale is recommended.

Table 4.6 EI Items with Changes Larger Than 0.1 Absolute Values in Category-Level Item Difficulty Between 5-Category and 4-Category Accuracy Scales

Item	Boundary thresholds							Change, Δb		
	5-category scale				4-category scale			(ΔSE)		
	$b_{1,\text{ 5-cat}}$	$b_{2,\text{ 5-cat}}$	$b_{3,\text{ 5-cat}}$	$b_{4,\text{ 5-cat}}$	$b_{1,\text{ 4-cat}}$	$b_{2,\text{ 4-cat}}$	$b_{3,\text{ 4-cat}}$	$b_{1,\text{ 4-cat}}$ $- b_{2,\text{ 5-cat}}$	$b_{2,\text{ 4-cat}}$ $- b_{3,\text{ 5-cat}}$	$b_{3,\text{ 4-cat}}$ $- b_{4,\text{ 5-cat}}$
Q10	-1.34	1.08	2.38	3.78	1.12	2.52	4.04	0.04	0.14*	0.26*
	(0.18)	(0.16)	(0.31)	(0.70)	(0.18)	(0.39)	(0.84)	(0.02)	(0.08)	(0.14)
Q11	-3.74	0.14	3.91	4.56	0.15	3.67	4.27	0.01	-0.24*	-0.29*
	(0.74)	(0.18)	(0.77)	(0.94)	(0.17)	(0.71)	(0.87)	(-0.01)	(-0.06)	(-0.07)
Q13	-1.35	0.34	2.05	2.33	0.33	1.90	2.14	-0.01	-0.15*	-0.19*
	(0.16)	(0.12)	(0.23)	(0.27)	(0.11)	(0.21)	(0.24)	(-0.01)	(-0.02)	(-0.03)
Q23	-3.06	-0.11	2.14	3.46	-0.11	2.21	3.57	0	0.07	0.11*
	(0.47)	(0.14)	(0.31)	(0.54)	(0.14)	(0.33)	(0.59)	(0)	(0.02)	(0.05)

Note. $b_{1,\text{ 5-cat}}$ = the boundary between the item score 0 and 1 or higher; $b_{2,\text{ 5-cat}}$, $b_{1,\text{ 4-cat}}$ = the boundary between the item score 1 or lower and 2 or higher; $b_{3,\text{ 5-cat}}$, $b_{2,\text{ 4-cat}}$ = the boundary between the item score 2 or lower and 3 or higher; $b_{4,\text{ 5-cat}}$, $b_{3,\text{ 4-cat}}$ = the boundary between the item score 3 or lower and 4;* Δ(b) > |0.1|.

Table 4.7 Selected Descriptive Statistics and Item Parameters of the Four Items on the 5-Category Accuracy Scale

Item	No of responses (percent)			Mean	Difficulty	Discrimination
	Category 0	Category 1	Category 4		(b_{overall})	(Change, Δ(a)*)
Q10	33 (17.10%)	117 (60.62%)	1 (0.52%)	1.10	1.48	-0.17 (= 1.66 – 1.77)
Q11	8 (4.15%)	95 (49.22%)	4 (2.07%)	1.48	1.22	0.08 (= 1.02 – 0.94)
Q13	30 (14.85%)	94 (46.53%)	8 (3.96%)	1.34	0.84	0.49 (= 2.64 – 2.15)
Q23	6 (2.97%)	91 (45.05%)	4 (1.98%)	1.60	0.61	-0.07 (= 1.32 – 1.39)

Note. * Change in item discrimination from the 5-category to 4-category scale, Δ(a) = $a_{\text{4-category}} – a_{\text{5-category}}$

Item Fit

Item fit analyses using S-X^2 (Orlando & Thissen, 2000) resulted in favorable outcome for the 4-category scale. The number of flagged items decreased when shifting from the 5-category to the 4-cagegory scale (see Table 4.8). At the conventional significance level (p < 0.05), six items (i.e., Item 2, Item 12, Item 15, Item 19, Item 23, and Item 33) were flagged on the 5-category scale. The use of the 4-category scale improved the fit of the misfitting items, except for Item 33, which left two items flagged (i.e., Item 15, Item 33). No items on either scale were flagged after the Bonferroni correction (p=α/number of items=0.0042). Details are found in Appendix 4.8.

Table 4.8 Statistically Flagged Items on the 5-Category and 4-Category EI Accuracy Scales with and without a Bonferroni Correction

Item	5-category scale				4-category scale				Change in item fit from 5-category to 4-category scale
	S-X^2	*df*	RMSEA	*p*	S-X^2	*df*	RMSEA	*p*	
Q2	45.00	26	0.06	0.012	37.68	27	0.05	0.083	improved
Q12	45.82	28	0.06	0.018	32.09	28	0.03	0.271	improved
Q15	52.18	33	0.05	0.018	48.36	31	0.05	0.024	(marginally) improved
Q19	60.37	38	0.05	0.012	37.23	36	0.01	0.412	improved
Q23	32.14	18	0.06	0.021	29.66	19	0.05	0.056	improved
Q33	50.51	28	0.06	0.006	52.07	29	0.06	0.005	(marginally) worsened
Number of flagged items									
p < 0.05				6				2	
p < 0.0042*				0				0	

Note. *significant with the Bonferroni correction

4.2.3 Person Statistics

Shifting from items to examinees, person fits were examined using Zh based on the *lz* statistic (Drasgow et al., 1985). Table 4.9 shows that responses on both scales fit similar in general (See Appendix 4.9 for details). The total number of persons with misfit (i.e., Zh > |± 2.0|) was 18 (2.31%) on the 5-categroy scale and 19 (2.44%) on the 4-category scale. Nine examinees were flagged on both scales, which means nine examinees' responses made improvement in their fit, but ten examinees response patterns were newly flagged on the 4-category scale. All misfitting response patterns except for one (i.e., ID F3-37) were underfit (i.e., Zh < -2.0), meaning that these individuals' responses were unexpected. Misfits were mostly mild, showing a Zh value smaller than |±3.0|. Only two examinees, ID F2-45 (Zh = -3.20) and ID F3-22 (Zh = -3.91), on the 5-category scale and one examinee, ID F2-62 (Zh = -3.02), on the 4-category scale was flagged, with their Zh values being larger than |±3.0|. One examinee's pattern (i.e., F3-37) was marginally overfit, which lacked variation on both scales, with its Zh values 0.23 and 0.24 on the 5- and 4-category scales, respectively. The person fit analyses conclude that the 4-category scale has the similar number of misfitting response patterns. Because the person fit statistics are comparable between the two scales, the more parsimonious scale is recommended.

Table 4.9 Person Misfits Measured on the 5-Category and 4-Category EI Accuracy Scales

5-Category scale			4-Category scale		
Person ID	Zh (*lz*)	Sub-total (Form)	Person ID	Zh (*lz*)	Sub-total (Form)
Form 1, Q1 to Q12 (N=193)					
F1-59*	-2.69	4	F1-73	-2.92	5
F1-73	-2.67		F1-40*	-2.52	
F1-65	-2.32		F1-65	-2.42	
F1-137	-2.16		F1-137	-2.34	
			F1-80*	-2.08	
Form 2, Q13 to Q24 (N=202)					
F2-45*	-3.20	4	F2-62	-3.02	5
F2-62	-2.96		F2-21*	-2.45	
F2-22*	-2.43		F2-20*	-2.24	
F2-16*	2.01		F2-132*	-2.22	
			F2-72*	-2.03	
Form 3, Q25 to Q36 (N=204)					
F3-22	-3.91	9	F3-63	-2.68	5
F3-36*	-2.88		F3-22	-2.55	
F3-204*	-2.70		F3-92	-2.32	
F3-151*	-2.69		F3-72*	-2.09	
F3-63	-2.49		F3-37	2.04	
F3-92	-2.21				
F3-20*	-2.03				
F3-111*	-2.01				
F3-37	2.03				
Form 4, Q37 to Q48 (N=180)					
F4-3	-2.19	1	F4-3	-2.30	4
			F4-63*	-2.11	
			F4-118*	-2.06	
			F4-142*	-2.01	
Total, Q1 to Q48 (% out of N=779)					
Zh > | ± 2.0|		18 (2.31%)			19(2.44%)
Zh > | ± 3.0|		2 (0.26%)			1(0.13%)

Note. *items flagged on one scale only; Person IDs are presented in an order of the Zh value size.

To summarize, examination of measurement qualities at the test, item, and person level generally indicated that the 4-category scale performed comparably with the current 5-category scale, or slightly better. Two items with sufficient number of responses rated as score zero, Item 10 and Item 13, demonstrated mixed results, but the variations between the two scales were not substantial. Particularly considering measurement precision of the lowest thresholds of the 5-category scale and parsimoniousness, the 4-category scale appeared to be a better alternative to the 5-category scale for the current sample.

4.3 Adequacy of Differentiating Exact Repetition from Appropriate Paraphrase on the EI Accuracy Scale (RQ1.2)

This section will discuss the evaluation of the adequacy of differentiating the exact repetition and errorless paraphrase (hereafter paraphrase) categories on the 4-category EI accuracy scale in terms of two purposes. First, the inspection was to collect measurement evidence that supports the higher order of exact repetition (the highest category) than paraphrase (the second highest category). Second, the evaluation was to determine whether the differentiation optimally fits the examinee population of the current testing program whose range of L2 English proficiency is restricted. Measurement statistics of the items on the 4-category scale confirmed higher ability of exact repetition than paraphrase but lacked consistency in empirical support for the use of the two categories across the items for the current population.

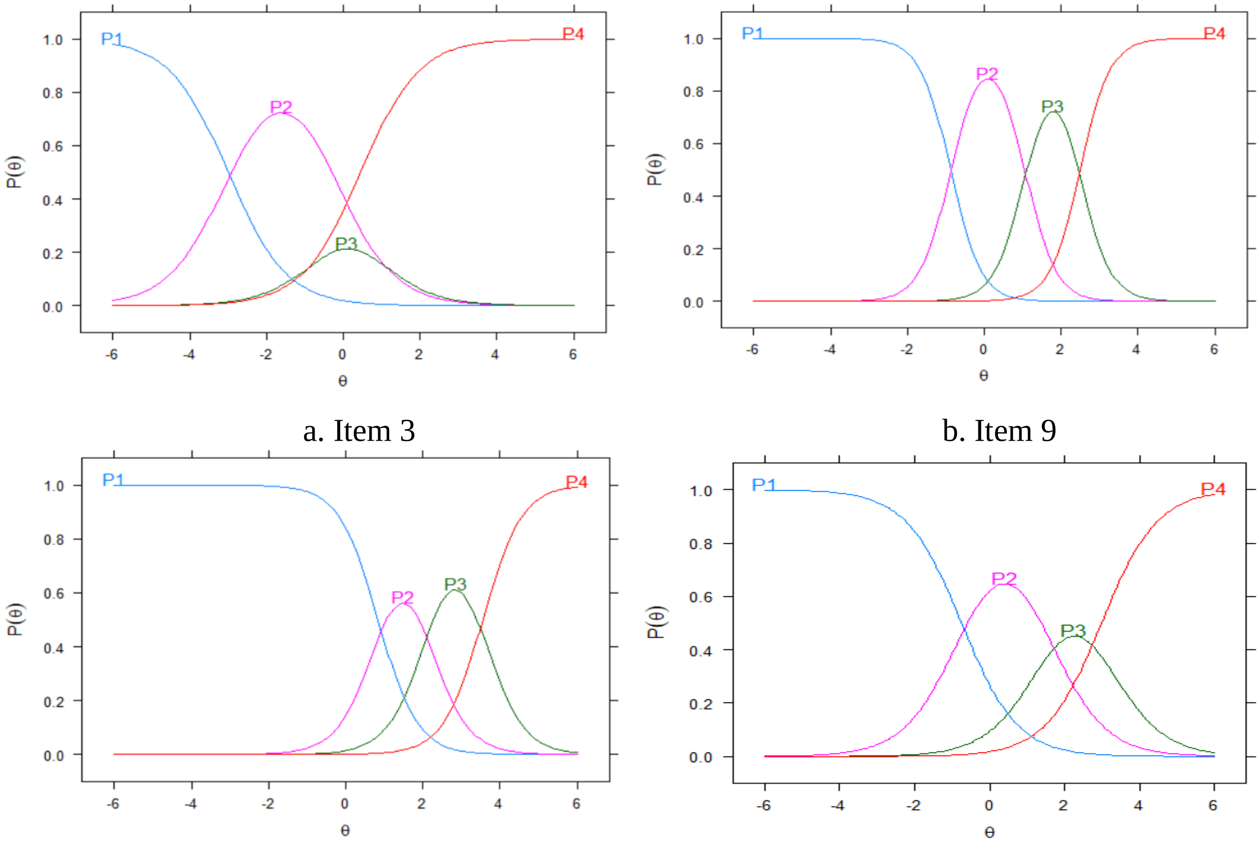
4.3.1 Higher Ordering of Exact Repetition over Paraphrasing

Overall, the measurement statistics of the GRM on the 4-category scale supported the theory on EI, the higher ordering of the exact repetition over paraphrasing. The item parameter statistics of the scores on the 4-category scale indicated that the cumulative category boundary threshold of the exact repetition (i.e., b_3; item score 4) was the highest in all 48 EI items (See Appendix 4.7 for details). The Category Characteristic Curves (CCCs) also displayed the highest ordering of the exact repetition category in all items. For example, each of the four items in Figure 4.3 in the next section positions P4 (i.e., the exact repetition category) on the higher value of the x-axis (i.e., the latent L2 English accuracy continuum), meaning responses rated as item score 4 has the highest latent trait of accuracy. In addition, the other three categories, P1 (i.e., the major error category), P2 (i.e., the minor error category), and P3 (i.e., the paraphrase category) were positioned in order, as the scale intended. Appendix 4.10 presents the CCCs for all items.

4.3.2 Empirical Support for the Adequacy of using the Paraphrase Category

Although the GRM model showed the higher ordering of the repetition than the paraphrase category on the 4-category scale, the results did not fully support differentiating the two highest categories for the current sample for practical benefits or precision issues. The distance between the accumulated threshold of the exact repetition and the threshold of the paraphrase category (i.e., b_3 - b_2) varied across the items, ranging from 0.14 (Item 30) to 2.58 (Item 28). Too small of a distance between the b_3 and b_2 parameters raised an issue about practical benefit of differentiating the two highest categories. In addition, the distance between the accumulated threshold of the exact repetition and paraphrasing categories (i.e., $b_3 – b_2$) was smaller than the distance between the paraphrasing and the minor error categories (i.e., $b_2 – b_1$) except for five items (i.e., Item 5, Item 8, Item 10, Item 21, and Item 28). The smaller distances between the two highest categories than between the lower categories also indicated marginal practical value of differentiating the exact repetition from the paraphrase categories.

The distances among the difficulty parameters for four example items are graphically illustrated in the CCCs of Figure 4.3 (See Appendix 4.10 for all items). The CCC of Item 3, *Last month we traveled to Chicago, which is the third largest city in the country*, in Figure 4.3 illustrates that all the surface under the curve line of the third category (denoted by P3) or the paraphrasing are shared with the adjacent categories. In other words, the P3 category, paraphrasing, is overlapping—practically redundant. For Item 22, *You can tell me what questions you have on the final project during my office hours*, the paraphrase category is more likely than the other options for the levels of L2 English accuracy approximately between the theta value of 1.8 and 2.4. However, the surface is small, which indicates a small difference, perhaps of little practical value. On the other hand, the CCCs of Item 9 and 10 indicated that the contribution of the paraphrase category is clear and sufficient, which provides empirical support of differentiating paraphrasing from exact repetition. Particularly, Item 10 demonstrated that unique information from the paraphrase category is larger than the minor error category. This pattern was found in the five items (i.e., Item 5, Item 8, Item 10, Item 21, and Item 28) in the current sample, which had greater distance between b_2 and b_3 than between b_1 and b_2, earlier.



a. Item 3

b. Item 9

c. Item 10

d. Item 22

Figure 4.3 Category Characteristic Curves for Item 3, 9, 10, and 22

Note. Item 3: *Last month we traveled to Chicago, which is the third largest city in the country*, Item 9: *The way that English classes are taught here might differ from the way in your country*, Item 10: *Purdue ranks second in (among) foreign student enrollment among (in) all public schools*, Item 22: *You can tell me what questions you have on the final project during my office hours*; x-axis: Position on the latent L2 English accuracy continuum; y-axis: probability of endorsement; P1: an irrelevant or incomprehensible response or response with major grammatical or semantic deviation, P2: a response with minor grammatical or semantic deviation; P3: errorless paraphrase with little meaning change; P4: exact repetition.

Table 4.10 presents the EI items classified by the amount of contribution from the paraphrase category. The classification was based on the graphic illustration and the total percentage of endorsement probability by the exact repetition category. In the current sample, more than the half of the total items (i.e., 27 out of 48 items) on the 4-category scale found little or no practical value of the paraphrase category, and eleven items with only a small value—the CCCs of these items are similar to Item 3 or 22 in Figure 4.3. Only ten items had a paraphrase category with reasonable or sufficient contribution. Even among the ten items, five items (i.e., Item 10, Item 21, Item 28, and Item 40) had an issue with measurement precision of the highest threshold in that the SE of their exact repetition category was higher than 4.0 due to the low frequency of the category. After all, the inspection of CCCs and SEs fully supported only six items (i.e., Item 5, Item 8, Item 9, Item 14, Item 47, and Item 48) for differentiating the two highest categories.

Table 4.10 Illustrated Practical Value of the Paraphrase Category on the 4-Cagegory EI Accuracy Scale

Illustrated Practical Value (IPV)	Graphical example in Figure 4.3	Number of items	Items
No or little	Item 3	27	Q1, Q2, Q3, Q4, Q7, Q11[a,b], Q12, Q13, Q15, Q16, Q17, Q18, Q19, Q20, Q25, Q29, Q30, Q31, Q32, Q35, Q37[b], Q38, Q39, Q41, Q42, Q44, Q46
Small	Item 22	11	Q6, Q22[b], Q23[b], Q24, Q26, Q27, Q33, Q34, Q36[b], Q43, Q45[b]
Reasonable or sufficient	Item 9 Item 10	10	Q5, Q8, Q9, Q10[b], Q14, Q21[b], Q28[b], Q40[b], Q47, Q48
Total		48	

Note. [a] an item of which the paraphrase category has SE value of 4.0 or higher; [b] an item of which the exact repetition category has SE value of 4.0 or higher.

All in all, the measurement statistics and CCCs of the GRM model on the 4-category scale confirmed the higher ordering of exact repetition than paraphrase in scoring EI. In other words, exactly repeating a prompt requires higher L2 English accuracy than paraphrasing the sentence. However, empirical support for the separate use of the two categories was neither prevalent nor consistent across the items, with more items unique information from the paraphrase category. . Plus, the independent use of the paraphrase category can raise a fairness issue—Is it fair to penalize appropriately paraphrasing, which is grammatically correct and has little meaning change from the prompt, compared to exact repetition. If paraphrased responses indicated lower lexical or grammatical complexity—for example, using *good* for *virtuous*, or *I mean* for *what I mean is that*—, differentiating paraphrase and exact repetition can be justified without a fairness issue. In other words, this kind of differentiation requires prompts to be lexically or syntactically difficult. The current data rarely included such cases. Thus, combining the exact repetition and paraphrase categories might be more reasonable for the current testing population. The findings called for two inquiries: 1) examining the use of the 3-category scale, compared with the 4-category scale

(Section 3.3.5), and 2) looking into item characteristics that are related to the adequacy of using both exact repetition and paraphrase categories for future item development (Section 3.3.4).

4.4 Item Characteristics Associated with the Adequacy of the Paraphrase Category (RQ1.3)

Since the practical value of the paraphrase category varied across the items, from little to substantial, the characteristics of the EI items on the 4-category scale were examined in association with the adequacy of the paraphrase category. The contribution from the paraphrase category was numerically calculated via the total probability of the endorsement (TPE) by the paraphrase category (P3), which is the sum of values of the fitted curve on a CCC for a given latent trait range, theta -4.0 to 4.0 for this analysis. Table 4.11 presents the selected correlations among the nine item (category) characteristics on the 4-category EI accuracy scale, including the TPE by the paraphrase category. To confirm the connection between the graphical evaluation of the paraphrase category based on CCCs, illustrated practical value (IPV) of the paraphrase category was coded as shown in Table 4.10 and added to the correlational analysis. Full description is found in Appendix 4.11.

Table 4.11 Correlations Among the (Category) Characteristics of EI Items on the 4-Cateogry Accuracy Scale

	Disc.	Difficulty (Threshold)			Distance	Responses Frequency (%)		TPE (P3)	Illustrated Practical Value (IPV, P3)
	(a)	(b_2)	(b_3)	($b_{overall}$)	($b_3 – b_2$)	(P3)	(P4)		
a	1								
b_2	-0.04	1							
b_3	-0.19	0.87***	1						
$b_{overall}$	-0.03	0.97***	0.93***	1					
$b_3 – b_2$	-0.32*	0.25	0.69***	0.42**	1				
P3 (%)	-0.18	-0.55***	-0.15	-0.42**	0.51***	1			
P4 (%)	-0.11	-0.79***	-0.89***	-0.86***	-0.61***	0.05	1		
TPE (P3)	-0.27	0.12	0.59***	0.30*	0.98***	0.63***	-0.56***	1	
IPV (P3)	-0.11	0.24	0.59***	0.39**	0.85***	0.40**	-0.58***	0.86***	1

Note. Disc. = item discrimination; b_2 = the accumulative threshold of the paraphrase category or higher; b_3 = the accumulative threshold of the exact repetition; $b_{overall}$ = overall item difficulty; Response frequency (%) = frequency of responses in the item category / total number of responses in the item × 100; P3 = the 3rd category (paraphrase); P4 = the 4th category (exact repetition); TPE (P3) = the total probability of endorsement by the third category (paraphrase) on a CCC; * p <0.05, ** p <0.01, *** p <0.001.

As expected, the TPE (P3) was almost perfectly and significantly correlated with the distance between the cumulative threshold of the paraphrase category (b_2) and that of the exact repetition (b_3) (r = 0.98, p <0.001). The perfect correlation means that the wider the latent trait range of the paraphrase category ($b_3 – b_2$) is the larger the contribution from the category is (TPE). This pattern is consistently observed, except some extreme cases where the fitted category curve is narrow and high-peaked while the adjacent curves are wide and low-peaked. IVP was also confirmed to strongly associated with the range ($b_3 – b_2$) and with TPE (P3), with the coefficients of 0.85 (p < 0.001) and 0.86 (p < 0.001), respectively.

The theta range of the paraphrase category (b_3 – b_2) was strongly correlated with the threshold of the exact repetition (b_3) (r = 0.69, p < 0.001) and moderately with overall difficulty (r = 0.42, p < 0.01) but not with the threshold of the paraphrase category (b_2) (r = 0.25, p = 0.09). This means the contribution of the paraphrase category is strongly associated with difficulty level of exact repetition rather than paraphrase. When a prompt is difficult to repeat exactly, examinees paraphrase the prompt. Similarly, the theta range of the paraphrase category (b_3 – b_2) was also significantly correlated with the percent of response frequency, negatively associated with exact repetition's (r = -0.61, p < 0.001) and positively with paraphrase's (r = 0.51, p < 0.001). These correlations indicated that (proportionally) the more responses in the paraphrase category and the fewer responses in the exact repetition category, the more difficult an item to repeat exactly, the greater the likelihood of paraphrase, and the greater the contribution of the paraphrase category.

The moderate but negative and significant correlation between the theta range of the paraphrase category (b_3 – b_2) with item discrimination (r = - 0.32, p = 0.03) is noteworthy and initially unexpected. The association, however, can perhaps be explained by sample characteristics. Generally, too easy or too difficult items tend to have lower discrimination power than items with medium difficulty. Because the current sample rated on the 4-category EI scale had only four easy items and the rest were either medium or difficult (See Table 4.5), items with medium difficulty, namely, less difficult items, tended to have higher discrimination and a lower threshold for exact repetition (b_3) and thus smaller contribution of the paraphrase category. Thus, the negative association should be carefully interpreted considering the item difficulty of the sample rather than being taken as a negative relation at face value or generalized. Cross validation would be needed.

Due to the high correlations among the three variables, the correlation pattern of the theta range (b_3 – b_2) with the item/category characteristics was similarly found with TPE (P3) and IPV

(P3), with magnitudes involving TPE or IPV being slightly smaller across the pairs, except for one case—response frequency (%) of the paraphrase category was slightly more strongly correlated with PTE (r = 0.63, p < 0.001) than with the theta range (r = 0.51, p < 0.001).

To conclude, among several item characteristics, item difficulty of the exact repetition category was most strongly associated with the practical adequacy of the paraphrase category. The unique information from the paraphrase category was sufficient, in other words, the ability levels at which the paraphrase category had the highest likelihood of endorsement are large enough, only when the item was difficult to exactly repeat. The findings will be discussed in relation with item writing and scoring in the Discussion section (Section 4.6).

4.5 Comparison of the 4-Category versus 3-Category EI Accuracy Scale (RQ1.4)

As shown in Section 3.3.3, empirical support from the GRM model conducted on the 4-category EI accuracy scale was insufficient for the use of the paraphrase category separately from the exact repetition category. As a follow-up, the 3-category GRM model, which collapsed the paraphrase category into the exact repetition category, was compared with the 4-category GRM model on test, item, and person levels.

IRT Assumptions

Prior to the comparison, scores from the 3-category scale were examined for the IRT assumptions of the unidimensionality and local independence, both of which were satisfied (See Appendix 4.3 and 3.4). The statistically flagged pair on the 4-category scale, Item 22 and Item 13 (LD-X^2 = 12.01) was not flagged on the 3-category scale anymore with its value decreased to 4.55. The improvement in the local independence increases statistical stability and confidence in the measurement on the 3-category scale.

4.5.1 Test Statistics

To compare the model fit of the GRM models on the 4- category and 3- category accuracy scales, several fit indices were examined. The overall relationships between items and the latent trait (i.e., L2 English accuracy) were compared using the sums of factor loadings and variance explained by each model. Scale information and conditional reliability were also examined.

Model Fit

Appendix 4.5 shows that Form 1, 2, and 3 (i.e., Item 1 to Item 36) fit well on both scales in terms of all statistics, showing non-significant M_2 statistics, RMSEA values of 0.01 or lower, and perfect TLI and CFI values. Form 4 (i.e., item 37 to 48) demonstrated slightly better performance on the 3-category when considering the cut-offs. On the 4-category, technically speaking, the *p*-value of Form 4 was significant (*p* = 0.0496) although the goodness of fit was acceptable to good (RMSEA = 0.07, SRMR = 0.06, CFI = 0.97, TLI = 0.96). The fit improved on the 3-catgory, displaying a clearly insignificant *p*-value (*p* = 0.11) and good fit (RMSEA = 0.04, SRMR = 0.05 CFI=0.99, TLI=0.99). Information criteria indices also demonstrated a better fit for the GRM models on the 3-category scale in that the AIC and BIC values were decreased across the four forms, with a range of difference from 728.27 (Form 4) and 848.16 (Form 2) in AIC values, and from 766.60 (Form 4) to 887.86 (Form 2) in BIC values.

Overall Factor Loadings

Similarly, overall performance of exploratory factor analysis was comparable between the two scales or slightly better on the 3-category scale (See Appendix 4.6). The sums of factor loadings and the proportion of variance explained slightly went up across all four forms with the maximum increase of 0.14 in the factor loadings (Form 2 and Form 4) and 1.2% in the variance proportion (Form 2). These little or slightly favorable changes on the 3-category scale means that collapsing the paraphrase and exact repetition categories does not weaken the relationships between items and the latent variable (L2 English accuracy). The comparable (or slightly better) magnitude of the relationships (i.e., sums of factor loadings, variance explained) suggests that the more parsimonious scale can be a good alternative, in line with the suggestion from the analysis of CCCs on the 4-category scale.

Test Information and Conditional Reliability

Figure 4.4 portrays the test information and conditional reliability of EI accuracy on the 4- and 3-category scales. The dotted and solid lines stand for the estimates on the 4- and 3-category scales, respectively. The scale information on the 3-category scale (y-axis) was reduced than on the 4-category scale for the higher theta ranges (x-axis) across the forms, which was expected because the two highest categories, the exact repetition and paraphrase categories, were combined. The theta levels at which the 4-cateogry scale starts providing more information varied across the forms. The lowest theta level was approximately -0.1 (Form 1) and the highest was 0.5 (Form 3 and Form 4), which the solid vertical lines mark. The information loss was maximized approximately at the theta level between 2.0 and 2.5 in general. However, at around the theta level 1, which the dotted vertical lines mark, observed information losses were small in Form 3 and Form 4. The information losses were not great in Form 1 and Form 2, although larger than in the other forms. The information loss in Form 1 and Form 2 was mainly due to the four items (i.e., Item 5, Item 8, Item 9, and Item 14), which had the sufficient amount of information from the paraphrase category with acceptable precision of exact repetition and paraphrase categories of the 4-category scale. Because the estimates in the theta are given in a standard normal metric, the theta score 1 means 84 percentiles of the L2 English accuracy. Thus, it can be interpreted that scores on the 3-categroy scale generally yielded the comparable amount of information about the examinees of 84.2 percentile or lower to that of the 4-category scale. The comparability in information supports that the more parsimonious scale is a good alternative.

Figure 4.5 illustrates the conditional reliability of EI accuracy on the 4- and 3-category scales. The dotted and solid lines respectively stand for the estimates on the 4- and 3-category scales. The conditional reliability estimates (y-axis) also indicated that the 4-category scale is slightly more reliable for the higher theta ranges (x-axis), starting from approximately the theta level between zero to 0.5, marked by the solid vertical lines. The distances between the solid and dotted vertical lines, meaning the differences in conditional reliability between the two scales, were maximized approximately at the theta level 4.0. In all four forms, the conditional reliability for the theta score 2.0 was high, being around 0.8 or higher. Because theta score 2 corresponds to examinees of 97.8 percentiles or lower, it can be interpreted that the majority of the scores were more than acceptably precise on the 3-category scale. The single IRT reliability estimates of the EI accuracy on the 3-category scales were also high, ranging from 0.86 (Form 4) to 0.89 (Form 3), and the maximum difference between the two scales was 0.01. It can be concluded that the 3-category scale did not worsen the reliability of the scores across different theta levels, compared to the 4-cateogry, and thus can be a good alternative scale.

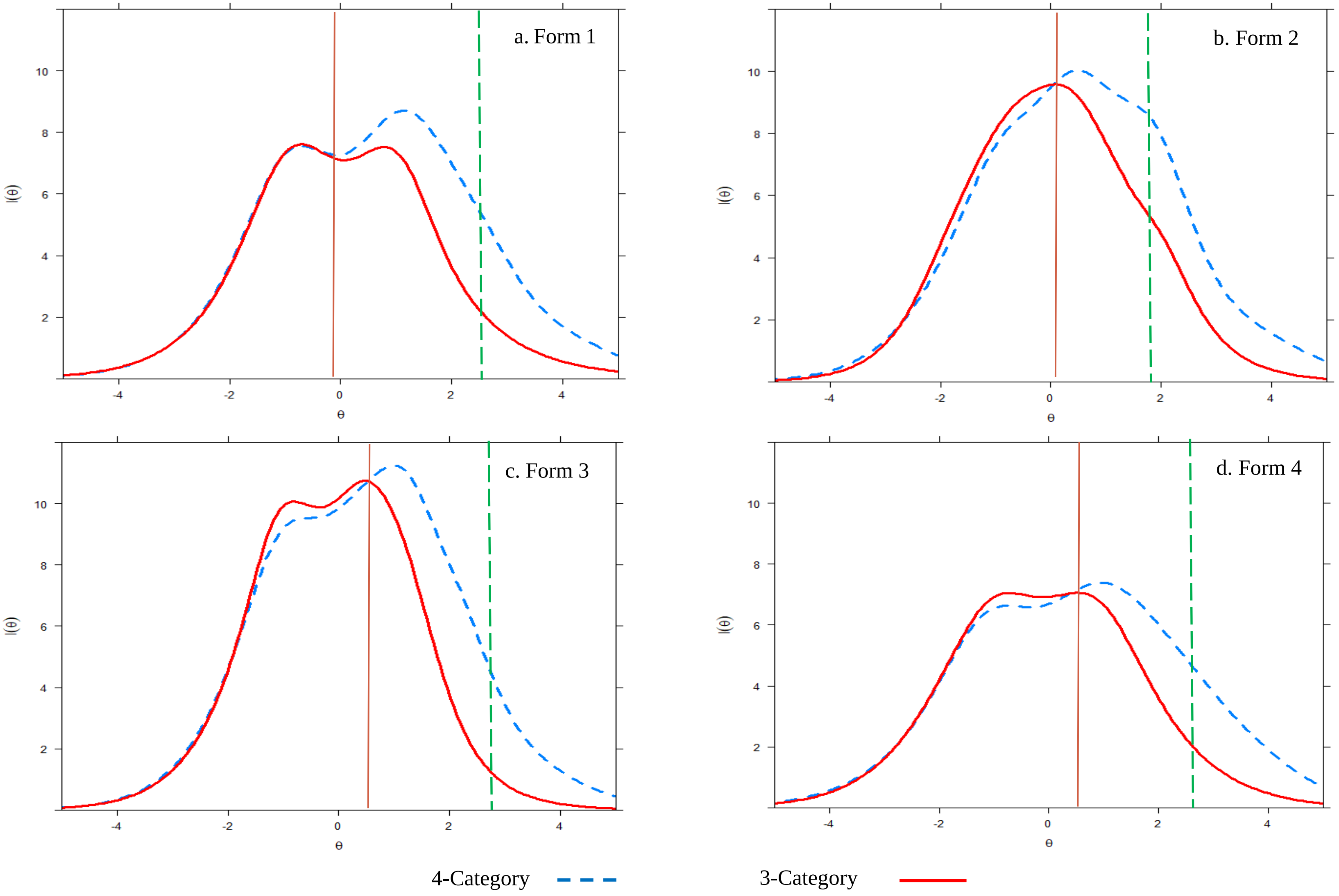


Figure 4.4 Test Information of the 4- and 3-Category EI Accuracy Scales

Notes. x-axis - ability level; y-axis - information

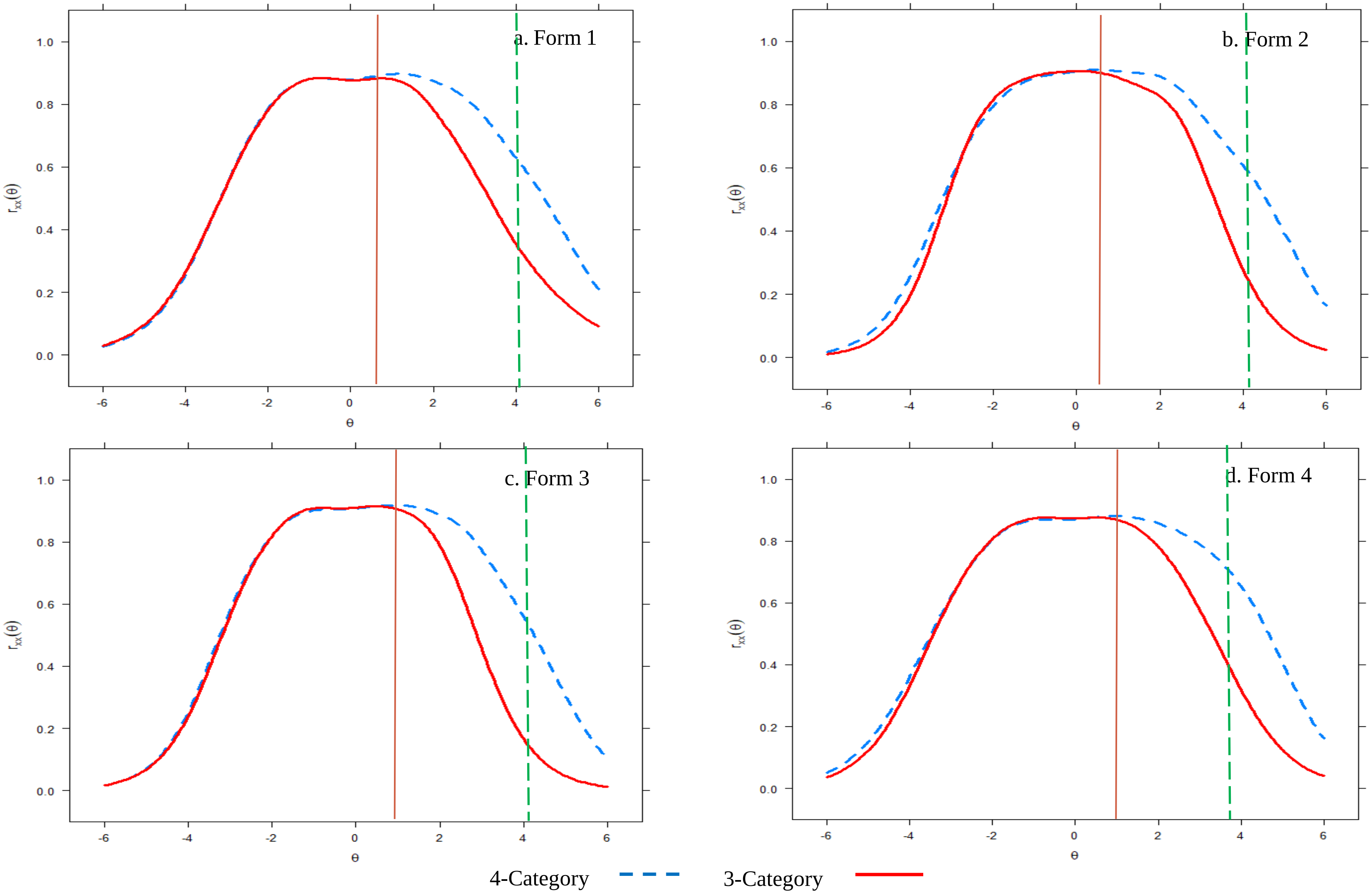


Figure 4.5 Conditional Reliability of the 4- and 3-Category EI Accuracy Scales

Notes. x-axis - ability level; y-axis – conditional reliability

4.5.2 Item Statistics

Item-level measurement statistics of the two GRM models were compared by examining item parameters and item fit. The item parameters examined are item discrimination and item difficulty coefficients at category levels and overall.

Item Discrimination

Overall, the use of the 3-category scale produced item discriminations comparable to those of the 4-category scale (See Appendix 4.7 for details). The correlation between the discriminations on the two scales was significant and strong (r = 0.94, $p < 0.0001$). The average increase across the 48 items on the 3-category was 0.04, with some items showing a non-marginal improvement. The paired dependent sample t-test indicated the changes of discrimination coefficients between the two scales were closely approaching the 95% significance level ($t = 1.66$, $p = 0.05$).

Table 4.12 shows that no item had (very) low discrimination on either the 4-category or 3-category scale. The range of the discrimination was larger on the 4-category scale, from 0.99 (Q11) to 2.83 (Q35), than on the 3-category scale, from 1.02 (Q11) to 2.64 (Q13). The distribution was similar on the lowest group, moderately discriminating items, in that the 3-category scale led to two more items only. The number of highly discriminating items went down from 24 on the 4-category scale to 18 on the 3-category scale, while the 3-category scale had four more very highly discriminating items (20 items) than the 4-category items (16 items).

Table 4.12 The Distribution of the Item Discrimination of the Scores Measured on the 3-Catgory and 4-Category EI Accuracy Scales

Discrimination (a)	No. of Items	
	4-category scale	3-category scale
Very low (a < 0.35)	0	0
Low (0.35 ≤ a < 0.65)	0	0
Moderate (0.65 ≤ b < 1.35)	8 (min. 1.02 (Q11))	10 (min. a=0.99, Q11)
High (1.35 ≤ b < 1.70)	24	18
Very High (b ≥ 1.70)	16 (max. 2.64 (Q13))	20 (max. a=2.83, Q35)
Total	48	48

Table 4.13 Changes in Item Discrimination (a) Estimates of the EI Scores Measured on 3-Category to 4-Category Accuracy Scales

Change in discrimination (Δ(a), $a_{4\text{-category}} - a_{3\text{-category}}$)	Frequency	Item (Δ(a), $a_{4\text{-category}} - a_{3\text{-category}}$)
Δ(a) > 0.2	4	Q18[a] (0.60), Q14[b] (0.41), Q35[c] (0.33), Q33[d] (0.26)
0.1 < Δ(a) ≤ 0.2	8	Q32 (0.20), Q43 (0.19), Q44 (0.17), Q9 (0.17), Q42 (0.15), Q4 (0.14), Q48 (0.12), Q39 (0.12)
-0.1 ≤ Δ(a) ≤ 0.1	31	The rest
-0.2 ≤ Δ(a) < -0.1	5	Q15 (-0.18), Q45 (-0.18), Q12 (-0.17), Q31 (-0.15) Q19 (-0.14)
Δ(a) < -0.2	0	N/A
Total	48	

Note. Q18[a] = *It looks like I only have morning classes this semester***,** Q14[b] = *Before you arrive on campus, you need to make sure that you have a place to live*, Q35[c] = *It can be very tough for foreign students to speak English on a daily (regular) basis*, Q33[d] = *The senior student was talking about his own story of (about) finding an apartment*

As shown in Table 4.13, examination of the item-level changes also indicated that overall item discrimination power was comparable between the two scales in that 31 items showed the change of ±1 or smaller. In approximately one third of items (i.e., twelve out of 48), discrimination was increased, with the maximum of 0.6 in Item 18 (*It looks like I only have morning classes this semester*). On the other hand, discrimination decreased in five items, with the maximum of -0.18 for Item 15.

To sum, item discriminations on the 3-category, the more parsimonious model/scale, were comparable to those of the 4-category scale, in general with slight improvement in some items and decrease in few items. Thus, 3-category scale can be considered as a good alternative to the 4-category scale.

Item Difficulty

Changes in item difficulty values were more noticeable than in item discriminations. Table 4.14 shows the item difficulty values on the 3-cateogry scale all went down in terms of overall difficulty (See Appendix 4.7 for details). The use of the 3-category scale resulted in more easy and medium-difficulty items, by the increase of four and eight in the number, respectively, while the number of (moderately) hard items decreased by 12 items. The 3-category scale did not have any very difficult items. According to Table 4.15, the overall difficulty decreased on the 3-cateogy scale by 0.62 on average. The differences ranged from -1.23 (in Item 28) to -0.23 (Item 30). The changes were significant from the results of a paired dependent sample t-test (t = -20.02, p < 0.00001). The decreases in overall difficulty (i.e., the average of all category-level difficulty estimates) were expected because the most difficult category was removed by combining the two highest categories of the 4-category scale into one on the 3-cateogry scale.

Table 4.14 The Distribution of Overall EI Item Difficulty Estimates ($b_{overall}$) Measured on the 5-Category and 4-Category Accuracy Scales

Difficulty	No. of Items		
	4-category scale	3-category scale	Difference
Very easy ($b_{overall} < -2.0$)	0	0	0
Easy ($-2.0 \leq b_{overall} < -0.5$)	4	8	4
Medium ($-0.5 \leq b_{overall} < 0.5$)	22	30	8
Moderately hard ($0.5 \leq b_{overall} < 1.0$)	8	5	-3
Hard ($1.0 \leq b_{overall} < 2.0$)	11	5	-6
Very Hard ($b_{overall} \geq 2.0$)	3	0	-3
Total	48	48	0

In contrast, Table 4.15 indicates that, in general, the category-level item difficulty estimates slightly increased at the threshold of Category 2 (i.e., minor semantic/grammatical errors) or higher, and the threshold of Category 3 (i.e., errorless paraphrase and exact repetition combined) (See Appendix 4.7 for details). The threshold of Category 2 (b_1) increased by 0.03 on average and the change ranged from -0.11 (Item 31) to 0.40 (Item 18). The changes between the scales were significant according to a paired dependent sample t-test (t = 2.50, p = 0.008). On the other hand, the average increase at the threshold of Category 3 (b_2) was slight, being 0.01, with the minimum of -0.08 (Item 43) to 0.21 (Item 45). Not surprisingly, the overall difference in b_2 was non-significant (t = 0.79, p = 0.22).

In addition to the categorical thresholds, Table 4.15 also presents distance between b_1 and b_2 (and distance between b2 and b3 for the 4-category scale) on each scale and the changes between the scales. On average, the distance on the 4-category (1.87) was slightly decreased on the 4-category (1.85) by 0.03, but the change was not significant (t = -1.49, p = 0.07). However, some items showed more than marginal changes, for example, Item 18, the distance ($b_2 – b_1$) of which was decreased by 0.4. The 3-category scale did not have either the threshold of Category 4 (b_3) or the distance between b_2 and b_3 because Category 4 was collapsed.

Table 4.15 Descriptive Statistics of Item Difficulty Coefficients on the 3- and 4-Category Scales and the Changes Between the Two Scales

Difficulty	4-category scale	3-category scale	Change (3-category – 4-category)
Overall ($b_{overall}$)			
Average	0.61	-0.01	-0.62
Range	3.45	3.24	1.00
Min., Max.	-0.75 (Q18), 2.70 (Q11)	-1.28 (Q48), 1.96 (Q11)	-1.23 (Q28), -0.23 (30)
Category 2 (b_1)			
Average	-0.97	-0.94	0.03
Range	3.93	3.70	0.51
Min., Max.	-2.81 (Q48), 1.12 (Q10)	-2.61 (Q48), 1.09 (Q10)	-0.11 (Q31), 0.40 (Q18)
Category 3 (b_2)			
Average	0.91	0.91	0.01 (≈ 0.913 – 0.906)
Range	4.06	4.16	0.29
Min., Max.	-0.39 (Q18), 3.67 (Q11)	-0.39 (Q18), 3.77 (Q11)	-0.08 (Q43), 0.21 (Q45)
Category 4 (b_3)			
Average	1.89	N/A	N/A
Range	3.99	N/A	N/A
Min., Max.	0.28 (Q32), 4.27 (Q11)	N/A	N/A
Distance ($b_2 – b_1$)			
Average	1.87	1.85	-0.03 (≈ 1.848 – 1.872)
Range	4.13	4.13	0.57
Min., Max.	0.61 (Q5), 3.52 (Q11)	0.62 (Q5), 3.62 (Q11)	-0.40 (Q18), 0.17 (Q15)
Distance ($b_3 – b_2$)			
Average	0.98	N/A	N/A
Range	2.44	N/A	N/A
Min., Max.	0.14 (Q30), 2.58 (Q28)	N/A	N/A

Note that although the changes in item difficulty coefficients vary in their size and significance, the correlations between item difficulty between the two scales were almost perfect: 0.97 ($p < 0.0001$) for overall item difficulty, 0.99 for b_1 ($p < 0.0001$), 1.0 ($p < 0.0001$) for b_2, and 0.97 ($p < 0.0001$) for the distance between b_1 and b_2. The very high correlations demonstrate that the changes occurred correspondingly across the items, meaning that the 3-category scale can serve as an alternative to the 4-category scale without distorting item difficulty drastically.

Relationships between Item Discrimination and Difficulty

To examine item characteristics associated with the changes of discrimination between the 4- and 3-category scales, correlations were calculated and presented in Table 4.16. Generally, the changes in item discriminations had weak and negative correlations (r = -0.28 or smaller, *p* = 0.05) with item difficulty coefficients on the 4-category scale (i.e., b_1, b_2, b_3, and $b_{overall}$). This means that discrimination coefficients of easier items on the 4-category scale slightly tended to be decreased on the 3-category more than difficult items. The associations were closely approaching the significance level (*p* = 0.05), so the small sample size (N = 48 items) might have led to lack of power.

Table 4.16 Correlation (r) Between the Changes of Item Discrimination ($a_{3\text{-category}} – a_{4\text{-category}}$) and Item Difficulty

	Coefficient r	(*p* value)
4-category scale		
discrimination (a)	0.12	(0.40)
difficulty, category 2 (b_2)	-0.28	(0.05)
difficulty, category 3 (b_3)	-0.28	(0.05)
difficulty, category 4 (b_4)	-0.28	(0.05)
difficulty, overall ($b_{overall}$)	-0.28	(0.05)
theta range (b_2 - b_1)	0.01	(0.93)
theta range (b_3 - b_2)	-0.03	(0.86)
3-category scale		
discrimination (a)	0.47	(< 0.001)
difficulty, category 2 (b_2)	-0.20	(0.18)
difficulty, category 3 (b_3)	-0.28	(0.05)
difficulty, overall ($b_{overall}$)	-0.26	(0.08)
theta range (b_2 - b_1)	-0.19	(0.20)
difference (3-category – 4-cagetory)		
difficulty, overall	0.10	(0.51)
difficulty, category 1 (b_1)	0.78	(< 0.0001)
difficulty, category 2 (b_2)	-0.62	(< 0.0001)
theta range (b_2 - b_1)	-0.90	(< 0.0001)

On top of individual threshold of each category, their theta ranges were also examined. The analysis revealed that the theta ranges of any categories on the 3- or 4-category scales alone were not correlated to the changes of item discrimination. Since the distance between b_2 and b_3 was strongly associated with the practical value of the paraphrase category, it can be assumed that the practical value alone was not associated with the change of discrimination, either.

Interestingly, however, the differences in the theta range of Category 2 (i.e., distance between b_1 and b_2) between the two scales showed a very strong negative association with the changes of item discrimination (r = -0.9, p = < 0.0001). Figure 4.6 illustrates the association. In Figure 4.6, when the theta ranges (b_2 - b_1) became smaller on the 3-category scale compared to the counterparts of the 4-category scale, the discrimination of the items increased on the 3-category. Item 18, whose theta range of Category 2 decreased the most among the 48 items, and the item most improved its item discrimination. Similarly, Item 15 showed the most widened theta range of Category 2 and the largest decrease in its item discrimination on the 3-category scale. Comparing the item information curves of the two items between the 4-category and 3-category scales which Figure 4.7 illustrates, the information provided by Item 18 visibly increased while the information by Item 15 decreased to some degree. Not surprisingly, the changes are aligned with the changes in item discrimination, because item information depends on item discrimination.

In addition to the change in theta range between b_1 and b_2, the differences between each category theta values were also significantly correlated with the change of discrimination values. The larger the differences of the theta value in Category 1 (r = 0.78, p = < 0.0001) and the smaller the differences in Category 2 (r = -0.62, p = < 0.0001) between the two scales were, the larger the increases in discrimination values were. The associations make sense because the increased threshold of Category 1 and/or the decreased threshold of Category 2 reduce the distances between the thresholds of the two categories.

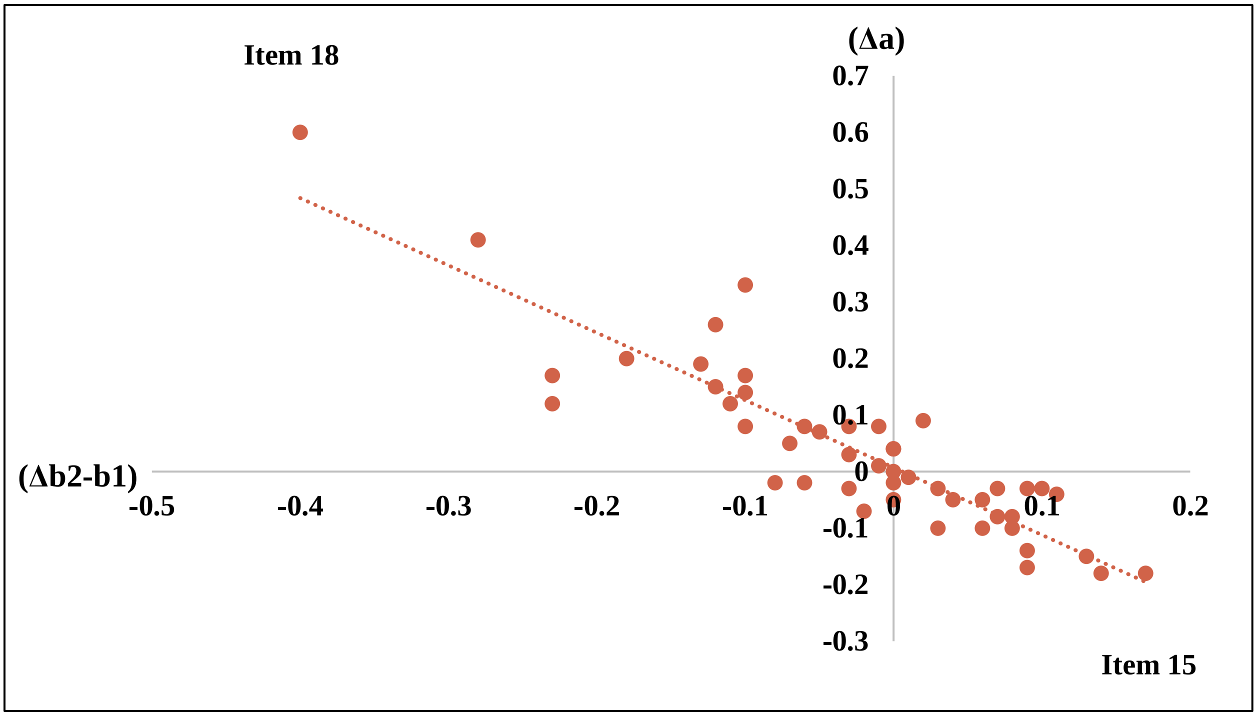
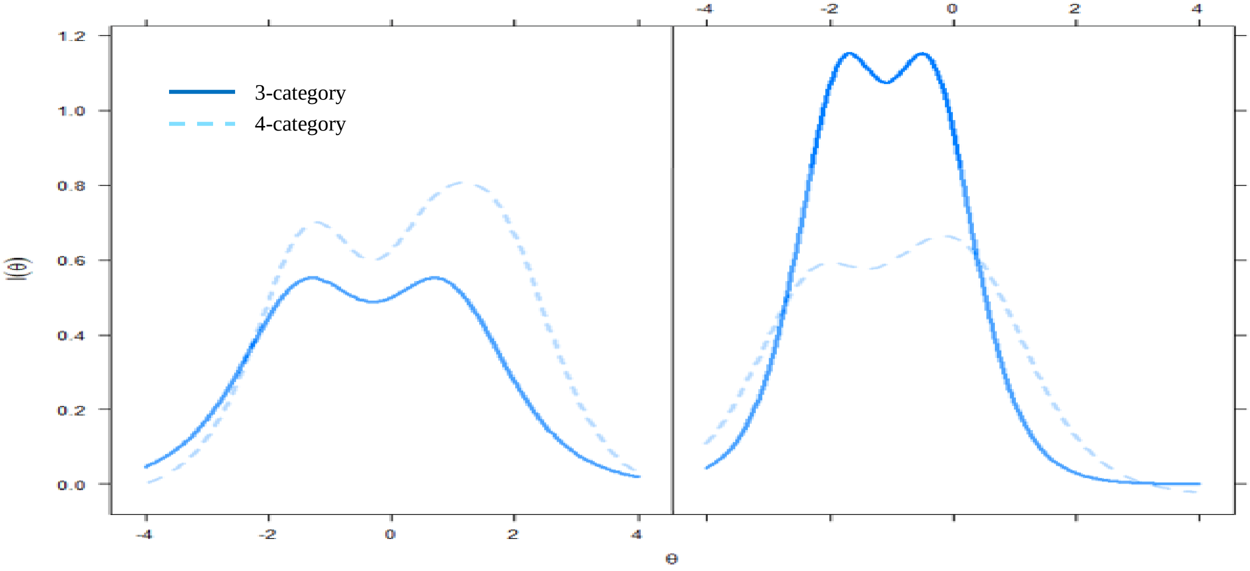


Figure 4.6 Relationship Between Changes in Item Discrimination (Δa) and Changes in the Theta Range of Category 2 (Δb_2-b_1) from the 4-Category Scale to 3-Category Scale



a. Item 15

b. Item 18

Figure 4.7 Item Information Curves and Category Characteristic Curves (CCC) (CCC) of Item 15 and Item 18 on the 4-Category and 3-Category Scales

Item Fit

Examination of the polytomous extension of $S\text{-}X^2$ (Orlando & Thissen, 2000) revealed that most items fit well on both 4-category and 3-category scales. There were only two statistically flagged items on the 4-category scale (i.e., Item 15, Item 33) and one item on the 3-category scales (i.e., Item 29) at the 95% significance level. Table 4.17 compares the item fit of the three items. On the 3-category scale, the items flagged on the 4-category scale, Item 15 ($p < 0.024$, RMSEA = 0.05) and Item 33 ($p < 0.005$, RMSEA = 0.06), improved their item fit, not being flagged any more. On the other hand, Item 29 was newly flagged on the 3-category scale ($p < 0.009$, RMSEA = 0.07). No item, however, was flagged with the Bonferroni correction ($p < 0.0042$) on either scale. The similar or slightly better overall performance in terms of item fit is aligned with the results from the analyses of the test and other item statistics, supporting that the more parsimonious scale can be a good alternative.

In order to see commonalities of the three flagged items, if any, information other than item fit was compared. Table 4.17 also presents item parameters and the size of the unique information from their paraphrase category of the three statistically flagged items. Few specific distinctions stood out comparing the three items with the non-flagged items although these items have some commonalities—the items ranked medium in their discrimination and difficulty within their forms.

Interestingly, however, Item 15 showed the largest decrease in the item discrimination from the 4-category to 3-category scale, but its item fit was improved on the 3-category scale. This seemingly contradictory finding make sense considering that previous research found no specific linear relationship between item discrimination and Orlando and Thissen' (2000) $S\text{-}X^2$ item fit statistic (Sinharay & Lu, 2008). The dilemma of decreased item discrimination and improved item fit invites close qualitative examination of Item 15 for the consistency of rating and examinee actual performance.

Another interesting observation is that Item 29, the only and newly flagged item on the 3-category scale had the paraphrase category with no practical value, which means that no ability levels were most likely endorse this category or no peak for the paraphrase category in the CCC. Meanwhile, Item 33, of which paraphrase category had some practical value, improved its item fit when the functioning category was collapsed. Although fully acknowledging that these two cases are an absolutely small number to infer any relationships, the cases might indicate that the practical benefit of the paraphrase category is independent of the changes in item fit. Since a few outliers can affect item fit, close qualitative examination of the misfitting items might reveal useful information on the fit of rubric, rater consistency, and actual examinee performance as well as item development.

Table 4.17 Changes in Item Statistics of Statistically Flagged Items from the 4-Category to 3-Category EI Accuracy Scales

Item	Scale	S-X^2	*df*	RMSEA	*P*	Item parameters (rank in Form)		Information
						Overall difficulty	Discrimination	(P3)
Q15	4-category	48.36	31	0.05	0.024*	0.35 (7th)	1.64 (6th)	little
	3-category	26.29	22	0.03	0.24	-0.30 (7th)	1.46 (8th)	N/A
Q29	4-category	22.28	17	0.04	0.17	0.13 (9th)	2.12 (4th)	no
	3-category	33.87	17	0.07	0.009*	-0.22 (6th)	2.19 (5th)	N/A
Q33	4-category	52.07	29	0.06	0.005*	0.31 (7th)	1.95 (6th)	small
	3-category	15.31	16	<0.01	0.50	-0.30 (7th)	2.21 (5th)	N/A
Number of flagged items (4-category)					2			
Number of flagged items (3-category)					1			

Notes. P3: *Statistically significant ($p < 0.05$); P3: Category 3; Q15 - *Joining a student club on campus is a great way to improve your social skills*; Q29 - *Regular workouts benefit the body as well as the mind*; Q33 - *The senior student was talking about his own story of finding an apartment*.

4.5.3 Person Statistics

Analyses of person fit slightly favored the 3-category scale. As shown in Table 4.18, overall, the number of examinees with an unexpected response pattern (i.e., Zh > | ± 2.0|) was smaller on the 3-category scale than the 4-category scale, slightly reduced from 19 to 15 examinees on the 3-category scale, which respectively accounted for only 2.43 % and 1.93 % of the total number of examinees. There were 12 examinees who were statistically flagged on both scales. All flagged responses were underfit except for one (ID F3-37), which means that the response patterns were unexpected, in other words, outliers. Only one flagged response on each scale showed a Zh value smaller than -3.0, ID F2-62 (Zh = -3.02) on the 4-category scale and ID F3-63 (Zh = -3.45). Thus, from a less conservative approach, only one response was flagged on each scale.

Table 4.18 Person Misfits Measured on the 4-Category and 3-Category EI Accuracy Scales

4-Category scale			3-Category scale		
Person ID	Zh (*lz*)	Sub-total (Form)	Person ID	Zh (*lz*)	Sub-total (Form)
Form 1, Q1 to Q12 (N=193)					
F1-73*	-2.92	5	F1-40*	-2.98	4
F1-40*	-2.52		F1-80*	-2.59	
F1-65	-2.42		F1-137*	-2.16	
F1-137*	-2.34		F1-73*	-2.04	
F1-80*	-2.08				
Form 2, Q13 to Q24 (N=202)					
F2-62	-3.02	5	F2-21*	-2.94	3
F2-21*	-2.45		F2-72*	-2.28	
F2-20	-2.24		F2-132*	-2.18	
F2-132*	-2.22				
F2-72*	-2.03				
Form 3, Q25 to Q36 (N=204)					
F3-63*	-2.68	5	F3-63*	-3.45	5
F3-22*	-2.55		F3-22*	-2.27	
F3-92	-2.32		F3-72*	-2.22	
F3-72*	-2.09		F3-71	-2.13	
F3-37	2.40		F3-74	-2.02	
Form 4, Q37 to Q48 (N=180)					
F4-3	-2.30	4	F4-118*	-2.51	3
F4-63*	-2.11		F4-63*	-2.15	
F4-118*	-2.06		F4-70	-2.03	
F4-142	-2.01				
Total, Q1 to Q48 (% out of N=779)					
Zh > | ± 2.0|		19 (2.43%)			15(1.93%)
Zh > | ± 3.0|		1 (0.13%)			1(0.13%)

Note. *items flagged on both scales; Person IDs are in order of the size of the absolute Zh values.

Interestingly, however, when the focus is put on the number of unexpected observed item scores that each examinee obtained, the 3-category scale notably outperformed the 4-category scale. For the current study, unexpected observed item scores are defined as the scores that show a difference of | ± 1.0| or larger than model-based expected item scores. The cutoff of | ± 1.0| is arbitrary but has a practical implication because each category is increased by 1 point on the current rating scale. Note that the criteria is not | ± 1| but | ± 1.0|. The number of unexpected responses considerably went up when using the criteria of | ± 1| is used, which requires model-based scores to be rounded without decimals and leads to classify differences between 0.5 and 0.9 as 1.0, discrepant scores. Thus, by using the criteria of | ± 1.0|, the study includes extremely unexpected cases only. As shown in Table 4.19, only 20 examinees had more unexpected item scores on the 3-cateogry scale than on the 4-category scale while 425 out of 779 examinees (57%) had a fewer number of unexpected item scores. The rest of the examinees (334 out of 779, 43%) demonstrated the same number of unexpected observations.

Table 4.19 Change in the Number of Unexpected Observed Items Scores Within an Examinee from the 4-Category to 3-Category EI Accuracy Scale

Change in the number of the unexpected responses* within an examinee (3-category – 4-category)	Number of Examinees
Examinees with more unexpected item scores on the 3-category scale (subtotal: 20)	
1	20
Examinees with no change (subtotal: 334)	
0	334
Examinees with fewer unexpected item scores on the 3-category scale (subtotal: 425)	
-1	219
-2	139
-3	50
-4	14
-5	1
-6	1
-7	1
Total	779

* Observed item scores showing a difference of |± 1.0| or larger from the model-based scores

Comparison of the unexpected scores obtained by examinees with person misfit revealed a similar pattern between the 4-cateogry and 3-category scales. Table 4.20 indicates that the number of unexpected responses by examinees with person misfit was much smaller on the 3-category scale than on the 4-category scale, decreasing from 79 to 48 items. This difference was mainly due to the unexpected scores endorsed by Category 4 (i.e., exact repetition) because out of the 31 unexpected items in Category 4, 28 items' scores were not considered unexpected anymore when the scores were converted to Category 3 on the 3-category scale. Decreases in the total number of unexpected items on the 3-category scale occurred across the misfitting examinees, except four who showed no change (i.e., ID F1-78, F2-21, F2-132, F3-37).

Table 4.20 Distribution of the Unexpected Item Scores Obtained by Examinees with Person Misfit on the 4-Category or 3-Category EI Accuracy Scales

ID	Flagged scale	Number of unexpected item scores			Exclusively unexpected score category (Item)	
		4-category	3-category	Difference	4-category	3-category
F1-40	both	2	1	1	1 (Q3)	-
F1-80	both	4	3	1	4 (Q4)	-
F1-137	both	5	3	2	4 (Q1, Q2)	-
F1-73	both	3	3	0	-	-
F1-65	4-category	4	2	2	4 (Q1, Q2)	-
F2-21	both	2	2	0	-	-
F2-72	both	6	4	2	4 (Q18, Q20)	-
F2-132	both	3	3	0	4 (Q19)	1 (Q22)
F2-62	4-category	4	2	2	4 (Q21, Q24)	-
F2-20	4-category	4	3	1	4 (Q15)	-
F3-63	both	5	3	2	4 (Q29, Q30)	-
F3-22	both	2	1	1	1 (Q32)	-
F3-72	both	3	1	2	4 (Q29, Q30)	-
F3-71	3-category	4	3	1	4 (Q30, Q32)	1 (Q28)
F3-74	3-category	3	2	1	1 (Q33)	-
F3-92	4-category	4	2	2	2 (Q31), 4 (Q35)	-
F3-37*	4-category	0	0	0	-	-
F4-118	both	4	3	1	4 (Q41, Q44)	3 (Q37)
F4-63	both	2	1	1	1 (Q44)	-
F4-70	3-category	6	4	2	4 (Q38, Q41, Q44)	3 (Q37)
F4-03	4-category	5	1	4	2 (Q44), 4 (Q40, Q46, Q48)	-
F4-142	4-category	4	1	3	1 (Q43), 4 (Q38, Q48)	-
Total		79	48	31	35	4

Note: * an examinee whose response pattern was overfitting

Overall, the fewer person misfit also supports the consideration for the use of the 3-category scale as a good alternative. Particularly, examination of the response patterns with aberrant items clearly favors the 3-category scale. Meanwhile, although misfitting responses were neither extremely severe nor prevalent across the items, the patterns and sources of misfits are worth investigating for test score validity and rating consistency.

To sum up, the 3-category scale generally demonstrated comparable or slightly better measurement performance at test, item, and person levels except for the scale information. The comparable performance of the 3-category scale is aligned with the observations from the CCCs of the 4-category scale in the previous section. The CCCs graphically illustrated that the majority of the items on the 4-category scale did not generate sufficient unique information from the paraphrase category. In the meantime, there were a few misfitting items and responses on the 3-category scale, although the number was reduced from the 4-category scale. The misfitting items and responses are examined further in Chapter 5 both quantitatively and qualitatively.

4.6 Discussion and Conclusion

Chapter 4 examined the adequacy of the use of categories in the EI accuracy scale for two purposes, first, to find the empirical justification for the higher order of exact repetition over errorless paraphrase, and second, to explore the number of scale categories for optimal measurement. First, noting that the lowest category of the current 5-cateogry scale (i.e., no, incomprehensible, or irrelevant response) was rarely chosen, the lowest category was collapsed into its adjacent category (i.e., response with major grammatical errors or major meaning difference). The comparison of the 5-category and 4-category scale favored the reduced scale in all areas examined except for conditional information and reliability at the lowest theta levels. The use of the 4-category scale led to higher measurement qualities at the test, item, and person level, particularly in measurement precision and parsimony. Since the examinees of the ACE-In EI scored 80 to 100 out of 120 on the TOEFL or a comparable score on the IELTS or DET, it was not surprising to observe that the extremely low frequency of the lowest category was prevalent across the items.

Combining the two lowest categories also appeared to ease the adverse effects of potential construct-irrelevance related to no response due to psychological or cognitive reasons other than examinee English proficiency. For example, there were a few examinees who performed reasonably well on most items but did not respond to the first item. This might have been due to test anxiety. Also, it is possible that individuals' personality increased no or irrelevant response since a few examinees chose to say *I don't fully remember.* or *Sorry, I forgot* instead of imitating at their best abilities.

Second, the higher L2 proficiency of the exact repetition over appropriate/errorless paraphrase was empirically supported, which backs up the 4-category scale. When carefully controlled, EI measures L2 proficiency rather than rote memory (Erlam, 2006; van Moere, 2021), while paraphrasing strengthens the construct-validity of EI as an oral proficiency measure (Yan et al., 2016). The higher order of exact repetition found in the IRT analyses adds empirical evidence for the theoretical founding of EI and increases confidence to use EI for testing and research purposes with the two categories included, in general. Meanwhile, it is also important to consider the limitations of the GRM model when interpreting the results. GRM assumes and imposes complete ordering of the categories. While the overall good model fit supported this assumption was met, some close category thresholds indicate either ill-separated categories or the lack of being truly ordered. Follow-up analysis based on Generalized Partial Credit Model, which allows partial ordering would clarify the distinction between these two possibilities induced by the GRM model.

Third, although there was evidence in support of the 4-level solution, the use of two separate categories for paraphrase and exact repetition did not gain full empirical support for the current population and items. In the substantial number of items, the paraphrase category did not provide sufficient unique information, which means little practical usefulness of the category. Also, the comparison of the measurement qualities between the 4-cateogry and 3-catgory scales favored the scale that combines the two categories in terms of measurement stability and preciseness. The appropriateness of separately using the two categories was positively associated with item difficulty. Using both exact repetition and paraphrase categories was adequate only when the threshold of the exact repetition category was appropriately high. In other words, ideal items are supposed to be easy enough for some examinees to repeat exactly, but simultaneously somewhat difficult for others to rather paraphrase. When the range of target language proficiency among the examinees is restricted, it is challenging to pinpoint the right levels of item difficulty. If the current items were tested on examinees of higher proficiency or L1 English speakers were included, the items would be evaluated with greater clarity. Thus, inclusion of advanced L2 speakers for comparison might be an option, as in previous research (e.g., Erlam, 2006). If the range of proficiency is restricted and broadening the range is not feasible, combining the two categories can lead to better measurement quality, which is the case for the current test taker population for the ACE-In EI.

One important aspect taken into consideration, however, the purpose of the test and target population to sharply discriminate. Collapsing the two highest categories loses information at the higher theta levels. Although the number of examinees affected by the category reduction is small, if discriminating the examinees at the highest proficiency levels is important, the reduced scale is not recommended. However, to reliably and precisely distinguish the proficiency levels associated with paraphrase and exact repetition separately, item revision is recommended rather than using the current items because many of the current items did not demonstrate measurement qualities that support the independent use of the two highest categories. Among the current items, there were five items that are appropriate for the 4-cateogry scale, which elicited sufficient practical value from the paraphrase category and demonstrated an acceptable range of precision for all four categories, including exact repetition: Item 5, Item 8, Item 9, Item 14, and Item 47. Information from the five items will be useful for item revision. Chapter 6 provides results from qualitative analysis for category adequacy, including the paraphrase category.

Lastly, despite the overall good fit and high/appropriate item difficulty and discrimination on the 3-category scale, there were a few misfitting items and responses. The misfitting items and responses were quantitatively and qualitatively examined, which will be discussed in Chapter 5 and Chapter 6.

CHAPTER 5. MISFITS ACROSS EXAMINEE ABILITY AND RESPONSE CATEGORY (RQ2)

This chapter presents the results of the quantitative examination of the misfitting items and responses (i.e., examinees) on the 4- and 3-category EI accuracy scales as well as the sources of the misfits. Misfitting examinees and items were descriptively analyzed for their patterns of score distribution across different ability levels and item score categories (Section 5.1 and 5.2) by answering the following questions:

RQ2.1 Where did the person misfits of the 4- and 3-category EI accuracy scales most occurred in terms of person ability and response category levels?

RQ2.2 Where did the item misfits of the 4- and 3-category EI accuracy scales most occurred in terms of person ability and response category levels?

The analyses of the responses with item and person misfits resulted in some associations of examinee ability and item categories with the fit (i.e., differences) between observed and IRT model-based item scores. Implications were made for item development and research (Section 5.3).

5.1 Person Misfit by Person Ability and Category Level (RQ2.1)

The GRM models run on the 4- and 3-category scales in Chapter 4 identified 22 examinees as misfits. The differences between observed item scores and model-based expected scores were examined for these examinees' responses by person ability and item score category to determine where the misfits occurred.

5.1.1 Person Misfit and Person Ability

Table 5.1 shows the response patterns of the 22 flagged examinees on either 3- and/or 4-category scales. Item responses that are likely to have made substantial contribution to the misfit of each response pattern (i.e., examinee) were marked with ^{a} and ^{b} for 3- and 4-category scales, respectively. In this study, the unexpected responses are operationally defined as observed item scores different from their corresponding IRT model-based scores by |±1.0| or more, not by |±1|.

Table 5.1 Response Patterns of Examinees with Person Misfit on the 3-Category (and 4-Category) EI Accuracy Scale

ID	*Zh*	Factor score	Flagged scale	Number of unexpected item scores	Observed Item Scores on the 3-category (and 4-category, if different)											
Form1					Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
F1-40	-2.98 (-2.52)	-1.17 (-1.16)	both	2 (3)	1	1	1 ^{b}	1	2	1	2	1	1	2 ^{a,b}	1	3 ^{a,b}
F1-80	-2.59 (-2.08)	0.38 (0.29)	both	4 (6)	3 ^{a,b}	2	1 ^{a,b}	3 (4 ^{b})	1 ^{a,b}	3	1 ^{a,b}	3 ^{b}	3	1	1	2
F1-137	-2.16 (-2.34)	1.09 (1.05)	both	3 (5)	3 (4 ^{b})	3 (4 ^{b})	3 (4)	3	3	3	3 (4)	1 ^{a,b}	2	3 ^{a,b}	2	1 ^{a,b}
F1-73	-2.04 (-2.92)	0.69 (0.36)	both	3 (3)	2	2	3	3	3	1 ^{a,b}	3	3	3	1	3 ^{a,b}	1 ^{a,b}
F1-65	-0.99 (-2.42)	1.13 (1.06)	4	2 (4)	3 (4 ^{b})	3 (4 ^{b})	3	3	3	3	3 (4)	3 (4)	2	1	3 ^{a,b}	1 ^{a,b}
Form2					Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24
F2-21	-2.94 (-2.45)	0.09 (0.10)	both	2 (2)	3 ^{a,b}	2	3	3	1	3 (4)	1 ^{a,b}	2	1	1	2	2
F2-72	-2.28 (-2.03)	-0.26 (-0.30)	both	4 (6)	1	3	1 ^{a,b}	1 ^{a,b}	1	3 (4 ^{b})	1 ^{a,b}	3 (4 ^{b})	2	1	1	3 ^{a,b}
F2-132	-2.18 (-2.22)	0.43 (0.30)	both	3 (3)	3 ^{a,b}	3	2	3	2	3	3 (4 ^{b})	1 ^{a,b}	2	1 ^{a}	1	2
F2-62	-0.97 (-3.02)	1.39 (1.11)	4	2 (4)	2	3	3	3	3	3	1 ^{a,b}	3 (4)	3 (4 ^{b})	3	1 ^{a,b}	3 (4 ^{b})
F2-20	-1.73 (-2.24)	0.52 (0.37)	4	3 (4)	2	3	3 (4 ^{b})	3	3	3	1 ^{a,b}	1 ^{a,b}	2	1 ^{a,b}	2	2
Form3					Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36
F3-63	-3.45 (-2.68)	-0.10 (-0.10)	both	3 (5)	2	2	1 ^{a,b}	2	3 (4 ^{b})	3 (4 ^{b})	1 ^{a,b}	2	1 ^{a,b}	1	3	2
F3-22	-2.27 (-2.55)	-1.12 (-1.13)	both	1 (2)	2	1	1	1	2	1	2	1 ^{b}	3 (4) ^{a,b}	1	1	1
F3-72	-2.22 (-2.09)	-0.20 (-0.27)	both	1 (3)	1	2	2	1	3 (4 ^{b})	3 (4 ^{b})	2	3	1 ^{a,b}	1	3	1
F3-71	-2.13 (-1.86)	0.01 (-0.04)	3	3 (4)	1	1	2	1 ^{a}	1 ^{a,b}	3 (4 ^{b})	3	3 (4 ^{b})	2	3 ^{a,b}	3	1
F3-74	-2.02 (-1.33)	-0.35 (-0.30)	3	2 (3)	2	2	2	2	3 (4) ^{a,b}	2	2	1 ^{a,b}	1 ^{b}	1	2	2
F3-92	-1.58 (-2.32)	0.84 (0.68)	4	2 (4)	2	1 ^{a,b}	2	1 ^{a,b}	3	3 (4)	2 ^{b}	3	3	3	3 (4 ^{b})	3
F3-37*	1.70 (2.40)	0.96 (1.02)	4	0 (0)	2	2	3	2	3 (4)	3 (4)	3 (4)	3 (4)	3	2	3 (4)	2
Form4					Q37	Q38	Q39	Q40	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48
F4-118	-2.51 (-2.06)	0.30 (0.27)	both	3 (4)	3 ^{a}	3	1	1 ^{a,b}	3 (4 ^{b})	2	3	3 (4 ^{b})	1	1 ^{a,b}	3	2
F4-63	-2.15 (-2.11)	-1.14 (-1.18)	both	1 (2)	1	1	1	1	1	3 (4) ^{a,b}	1	1 ^{b}	2	1	2	3
F4-70	-2.03 (-1.74)	0.26 (0.36)	3	4 (6)	3 ^{a}	3 (4 ^{b})	3 ^{a,b}	2	3 (4 ^{b})	1 ^{a,b}	2	3 (4 ^{b})	1	2	2	1 ^{a,b}
F4-03	-0.69 (-2.30)	0.85 (0.84)	4	1 (5)	2	1 ^{a,b}	2	3 (4 ^{b})	3	3	3	2 ^{b}	2	3 (4 ^{b})	2	3 (4 ^{b})
F4-142	-1.36 (-2.01)	-0.07 (-0.02)	4	1 (4)	2	3 (4 ^{b})	1	1	2	1 ^{a,b}	1 ^{b}	3	2	3	2	3 (4 ^{b})

Note. $^{a, b}$ unexpected observed item scores, which is different from the corresponding model-based expected scores by | ±1 | (i.e., top 10% the difference) on the 3-category ^{a} and 4-category scale ^{b} ;*an overfitting item

Table 5.1 also examinee factor scores and the total number of the unexpected items for each misfitting examinee.

Table 5.1 does not show any specific identifiable pattern across the 22 examinees of person misfits, in terms of their ability. Twelve examinees' ability levels were medium level, -0.5 to 0.5 on the 3-category scale, while seven examinees belong to the higher end, showing a lager theta value that 0.5. There were only three examinees whose ability level was below 0.5. The pattern was similar on the 4-category scale with two examinees shifting from the higher end to the middle ability group. Because the theta score of -0.5 to 0.5 covers approximately 38% of the population, the medium ability group accounted for the misfitting response patterns proportionally greater than the other ability groups. It is not surprising that the middle ability group is more vulnerable to misfits, which is generally the case in language tests. Also, the smaller number of misfitting examinees in the low proficiency level than the high proficiency level makes sense because it is fairly impossible to perform substantially better than one's ability level compared with the opposite case, where highly proficient examinees do not perform as expected. That is, several factors can negatively affect one's performance on one or more responses, for example, test anxiety, response tendency, or fatigue effects. However, factors that can lead to higher performance on EI than one's English proficiency is relatively limited other than items with very low discrimination, such as very easy items where examinees of low proficiency can obtain high scores or very difficult items where few examinees perform well.

5.1.2 Person Misfit and Item Score Category

The patterns of unexpected item scores in relation with item category levels were more obvious among the responses by examinees with person misfit. According to Table 5.1 and 5.2, the large number of unexpected responses in the exact repetition category obtained by the examinees with person misfit resulted in a great difference in the total number and pattern between the 3- and 4- scales, otherwise very similar. Excluding the exact repetition category, both scales showed the largest number of unexpected scores in the major error/deviation category, some cases on the paraphrase category, and only marginal or no unexpected responses in the minor error/deviation category.

Table 5.2 indicated the major error/deviation category accounted for 45.12% of the misfitting examinees' unexpected item scores on the 4-category scale, and 68% on the 3-category scale. All examinees with person misfit on either category scale had one or more unexpected responses to major error/deviation, except for one examinee with overfit (ID F3-37). Notably, 31 responses to exact repetition were unexpected. When the two highest categories were combined, however, the majority (i.e., 28 out of 31 responses), were not flagged, which was the major difference between the two scales. The minor error/deviation category showed the lowest, followed by the paraphrase category.

Table 5.2 The Number of Unexpected Item Scores among the Misfitting Examinees' Responses by Category

Scale	Number of unexpected item scores by category				Total
	1	2	3	4	
4-category	37 (45.12%)	3 (3.66%)	11 (13.41%)	31(37.80%)	82
3-category	34 (68.00%)	1 (2.00%)	15 (30.00%)	N/A	50
Sub-total	71 (53.79%)	4 (3.03%)	26 (19.70%)	31(23.48%)	132

Notes. Category 1 - major grammatical and semantic deviation or no/irrelevant/incomprehensible responses; Category 2 - minor grammatical errors or semantic deviation; Category 3 – errorless paraphrase; Category 4 – exact repetition.

In summary, the examination of distribution of examinees with person misfit on the two EI accuracy scales revealed that fewer examinees of low proficiency and the greater number of unexpected responses in the extreme categories, particularly the lowest category for both scales, and the highest category for the 4-category scale only. These findings might suggest some potential associations of person ability and score category levels, including their interaction, and invite an examination beyond the misfitting examinees' responses, which will be discussed in Chapter 7. Also, the large number of unexpected responses found in the lowest category than in the other categories call for qualitative examination of responses (and rater comments) to check how raters applied the rating criteria of the category, specifically regarding decision-making on the quality of grammatical errors and semantic deviation, whether they are major or minor. Chapter 6 will present the qualitative findings.

5.2 Item Misfit by Person Ability and Category Levels (RQ2.2)

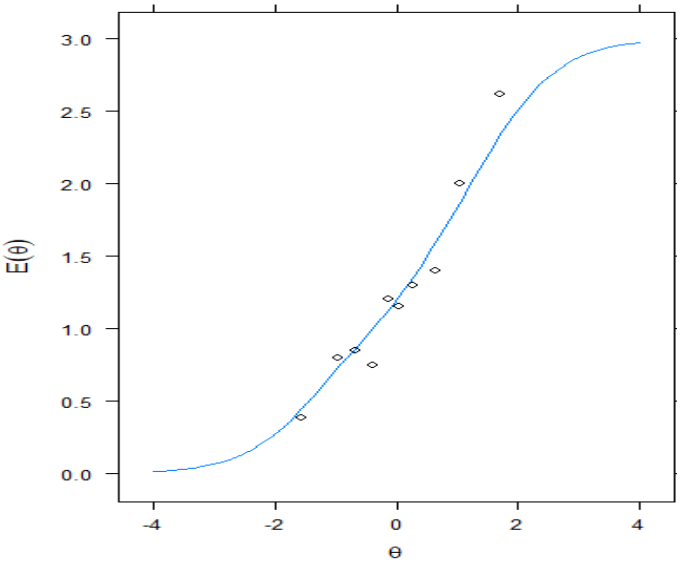
In Chapter 4, the GRM models identified three misfitting items—two on the 4-cateogry scale (i.e., Item 15, Item 33) and one on the 3-category scale (Item 29)—at the 5% significance level. To examine where the misfits occurred in the three items, the differences between the frequency of empirical/observed scores and expected scores from the GRM models were analyzed across examinee ability (ϴ) levels and item score categories.

5.2.1 Item Misfit by Person Ability

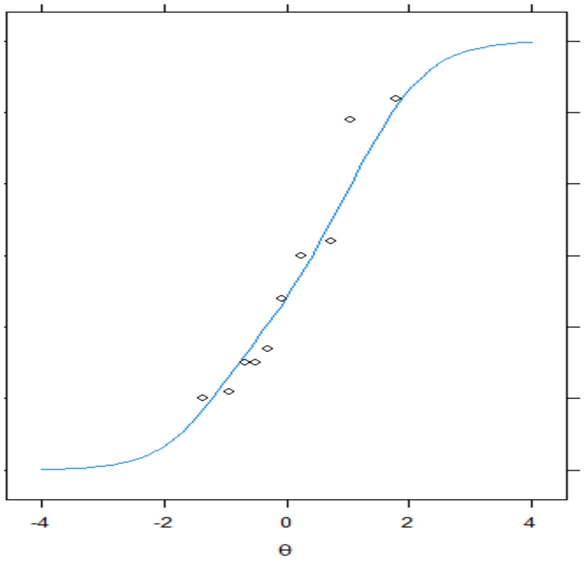
Figure 5.1 displays the fit between observed scoring and IRT model-based expected scoring in the three misfitting items across different person abilities. In each plot, the GRM-based expected scores is displayed as a curve and the empirical or observed scores, which collapses the polytomous item categories, is plotted with circles. The circles closer to the curves mean smaller differences between the observed and model-based scores. In Figure 5.1, there were some circles distanced from the curves, which indicates that the empirical and model-based scoring of the three items differ from each other to some degree, although not drastically. The graphical descriptions are aligned well with the item fit of the items. These three times were flagged at the 95% significance level—Item 15 (RMSEA = 0.05, *p* = 0.024), Item 33 (RMSEA = 0.06, *p* = 0.005), and Item 29 (RMSEA = 0.07, *p* = 0.009)—but their RMSEA values were borderline, and the items were not flagged with the Bonferroni correction (*p* = 0.0042). Figure 5.1 also revealed that the distances differ across the ability levels and the discrepancies were relatively larger at higher ability levels and smaller at the lowest levels to some degree with some noise.

The between-ability differences in fit between empirical data and the models are illustrated more clearly in Figure 5.2. Figure 5.2 plots the differences in frequencies between the empirical and model-based scores (not the scores themselves) by ability percentiles of the current sample. Because Figure 5.2 and Figure 5.1 employed the unstandardized differences, the distance between the circles and the curves in Figure 5.1 are more directly connected with the differences between the observed and expected frequencies in Figure 5.2 while Figure 5.3 plots the standardized residuals (i.e., z.residuals) based on z statistic. Both figures provide the maximum and minimum values only but all numerical values corresponding to Figure 5.2 and Figure 5.3 are presented in Appendix 5.2. (See Appendix 5.1 for the theta by category distributions). Figure 5.2 and Figure 5.3 suggest that there are some variations in the (standardized and unstandardized) differences between empirical and model-based frequencies at different ability levels, in general. However, the overall sizes of the differences varied between the three items, with Item 15 having the largest. Also, from the linear trendlines, the overall tendency, the larger differences at higher theta levels, was revealed. However, there were variations in the strengths of the relationship among the items. When the frequency differences were standardized as z.residuals (in Figure 5.3), the patterns changed slightly .

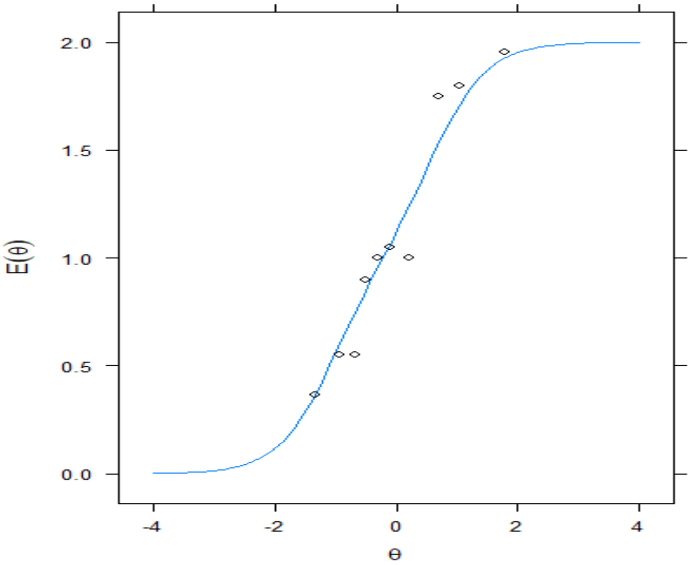
In summary, the analysis of the misfitting items revealed some potential that examinee ability levels are associated with the extent to which the observed item scores are different from the model-based item scores and the frequencies of the scores are. Also, the analysis suggested that the magnitude varied across the items.



a. Item 15 (4-Category Scale)



b. Item 33 (4-Category Scale)



c. Item 29 (3-Category Scale)

Figure 5.1 Empirical and Model-Based Frequency for Expected Scores by Ability in the Three Misfitting Items on the EI Accuracy Scales

Notes. Item 15 (the 4-category scale): *Joining a student club on campus is a great way to improve your social skills*; Item 33(the 4-category scale): *The senior student was talking about his own story of (about) finding an apartment*. Item 29 (the 3-category scale): *Regular workouts benefit the body as well as the mind*.

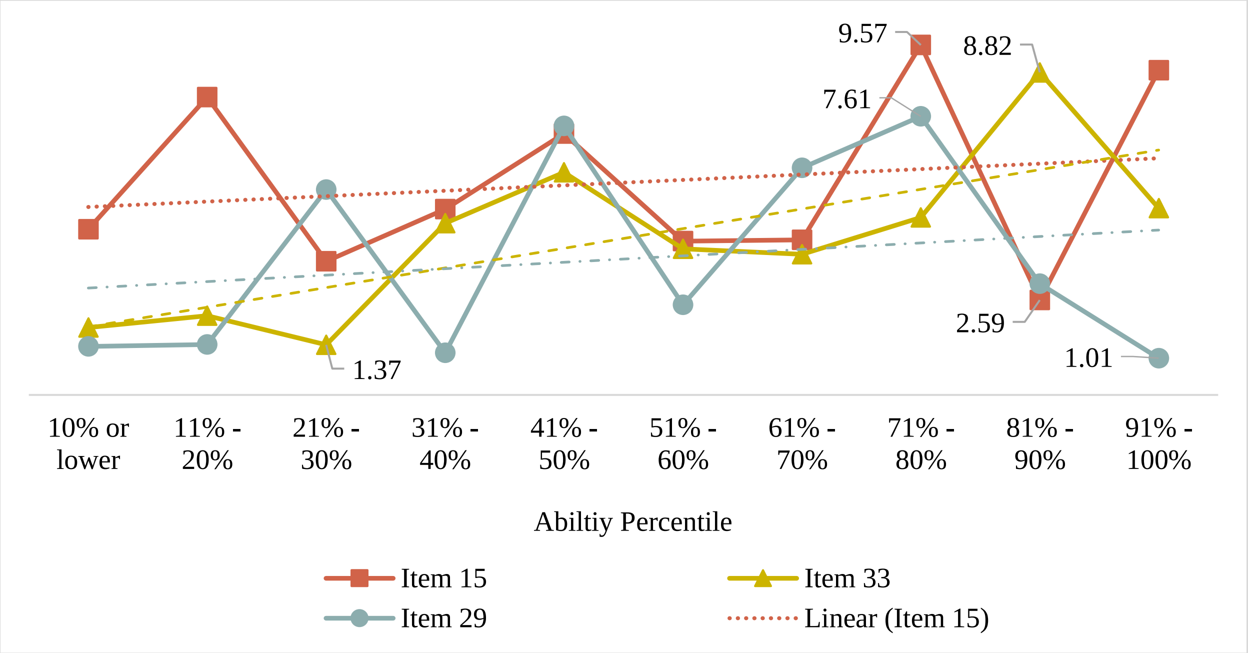


Figure 5.2 Differences between the Observed and the Expected Frequency (in Absolute Values) by Ability Percentiles

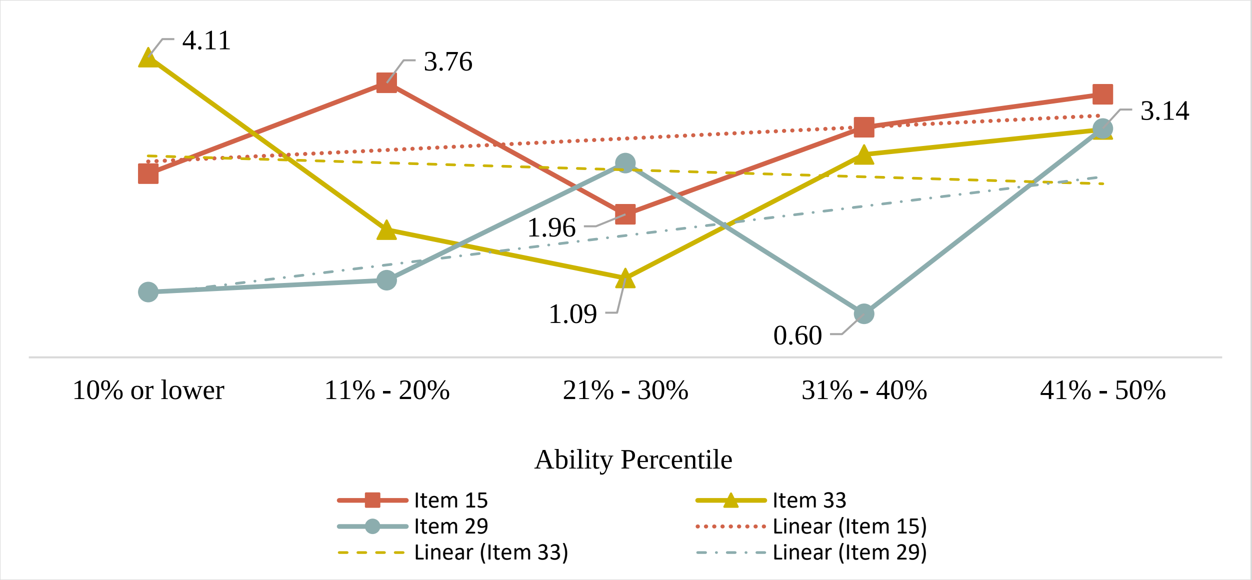
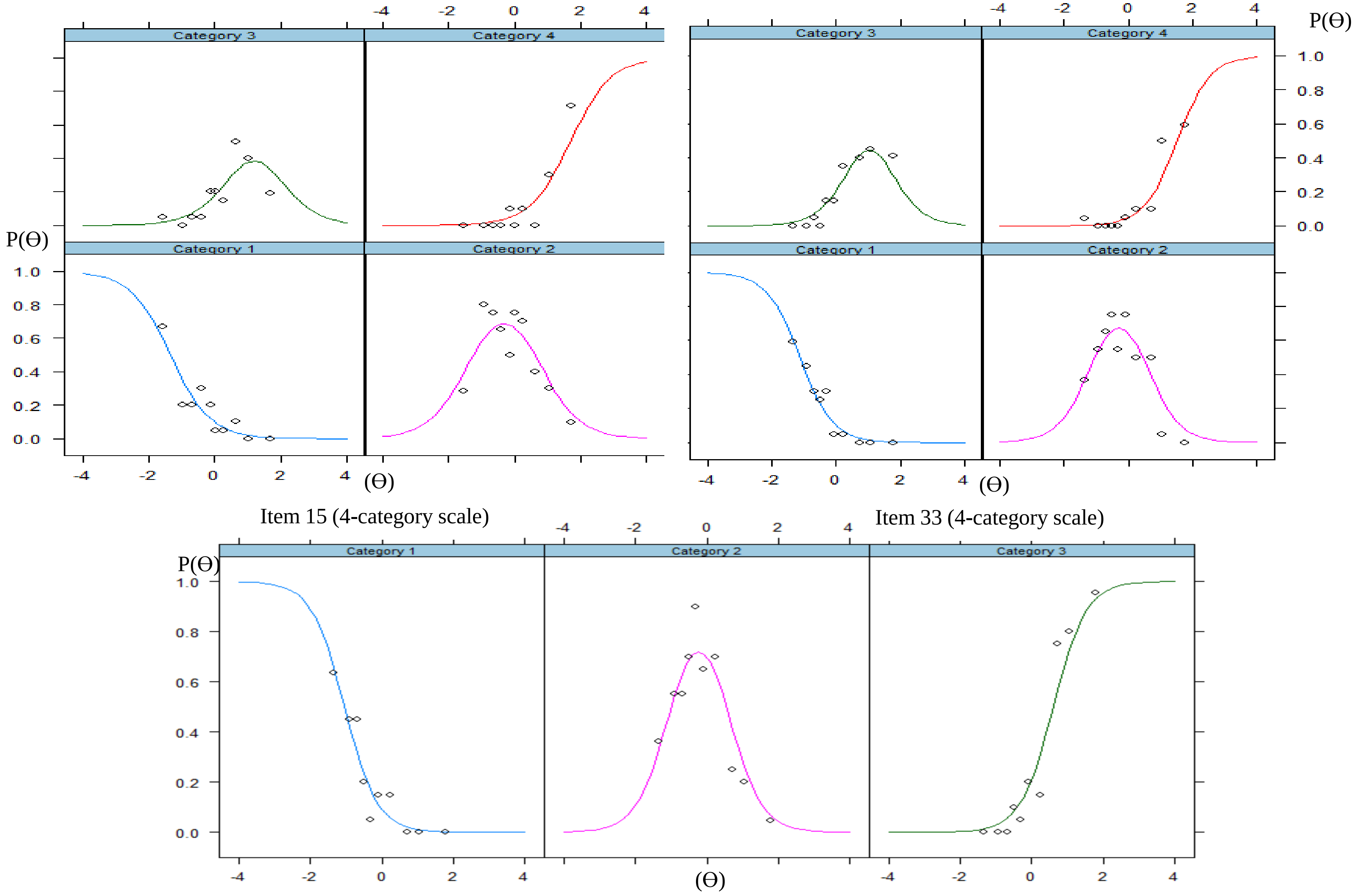


Figure 5.3 Z Statistic Residuals (in Absolute Values) of the Observed Frequency by Ability Percentile

5.2.2 Item Misfit by Response Category Levels

The potential association of the fit (between empirical scores and model-based expected scores) with the response categories was more apparent than the one with the ability levels. Figure 5.4 presents the fit between the data and GRM models by the response categories of the three misfitting items (i.e., Item 15, Item 33, and Item 29). The model-based CCCs are displayed as curves and the empirical CCCs are plotted with circles. Larger distances mean worse fit. Overall, variations in the fit were observed between the categories as well as between the items. Across the items, the empirical CCCs were most distanced from the model-based CCCs in Category 2, and most closely aligned in Category 1. The responses to Item 15 were more scattered than the two items, which means larger differences between the empirical and model-based scores.



Item 15 (4-category scale)

Item 33 (4-category scale)

Item 29 (3-category scale)

Figure 5.4 Empirical and Model-based CCCs for the Three Misfitting Items on the EI Accuracy Scales

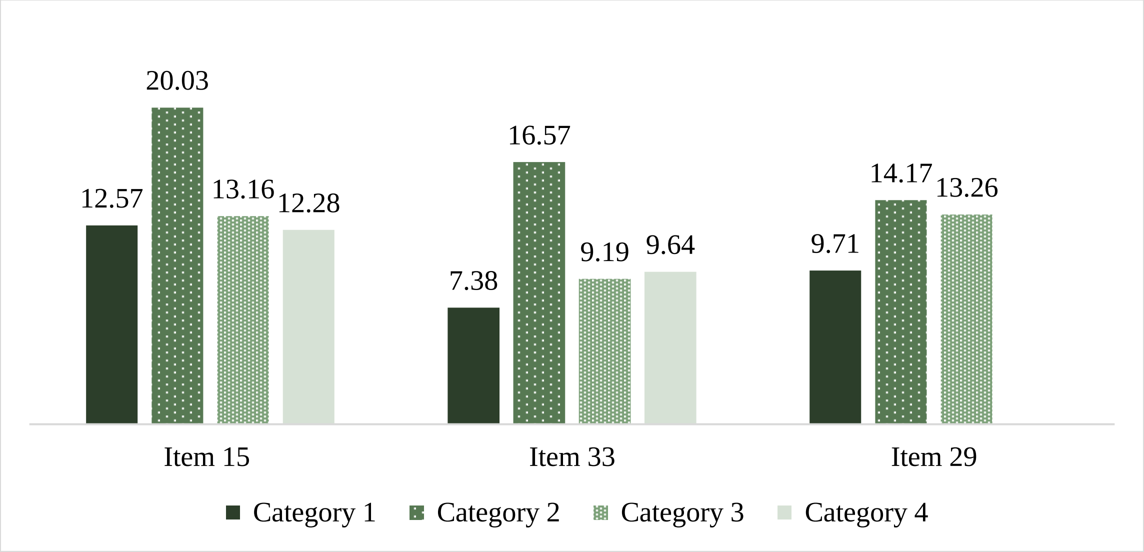


Figure 5.5 The Differences in the Frequencies of the Empirical and Model-Based Responses to the Three Misfitting Items on the EI Accuracy Scales by Response Categories

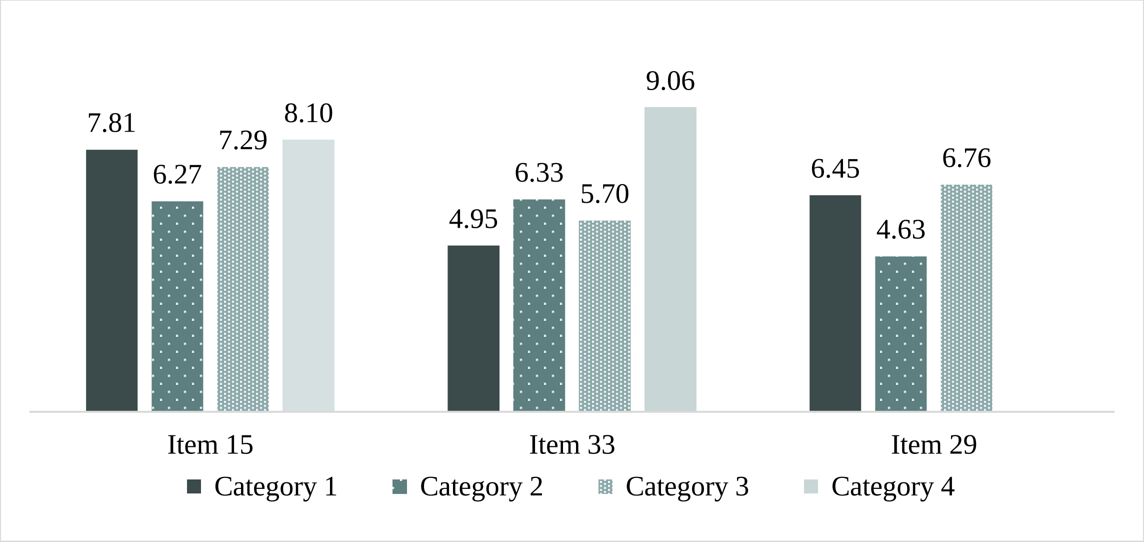


Figure 5.6 Z.Residuals of the Observed Scores of the Three Misfitting Items on the EI Accuracy Scales by Response Categories

The fit differences between response categories and the items were demonstrated more clearly in Figure 5.5. Figure 5.5 above shows the differences in the frequencies between empirical and expected scores by response categories in each item. Figure 5.6 illustrates the same information using z-residuals based on z statistic, not the unstandardized scores. The same information from the Figure 5.5 and Figure 5.6 is also presented in tables in Appendix 5.3.

As in the comparison of empirical and model-based CCCs, Figure 5.5 indicated the largest discrepancy in the observed and model-based frequencies in Category 2 and (approximately) the least in Category 1 across all three items. Also, the extent to which the discrepancy varied by category as well as the total amount of discrepancy was not consistent among the three items.

Interestingly, however, Figure 5.6 demonstrates when the differences were standardized, the pattern was different. Category 4 was found to be most discrepant for Item 15 and Item 33, which were calibrated on the 4-category scale. The trends in the other three categories were not consistent across the items, either, in that Category 2 showed the least differences in Item 15 and Item 29 while Category 1 was in Item 33.

To examine the potential reasons why Item 33 behaved differently from the other two items, the distribution of the responses (in Table 5.3) and parameters (in Table 5.4) were compared. In Table 5.4, When compared the number of responses among the three items, no considerable differences were observed.

Table 5.3 Number of Responses by Response Category of the Three Misfitting Items

Items	No. of Responses (Percent)				
	1	2	3	4	Total
4-category scale					
Q15	36 (17.82%)	105 (51.98%)	36 (17.82%)	25 (12.38%)	202 (100%)
Q33	41 (20.10%)	94 (46.08%)	40 (19.61%)	29 (14.22%)	204 (100%*)
3-category scale					
Q29	43 (21.08%)	99 (48.53%)	62 (30.39%)	N/A	204 (100%)

Note. *The total is not 100% due to the round up.

The item parameters presented in Table 5.4 did not directly point out notably different item behavior in Item 33 compared with Item 15 and 29. The overall difficulty, the threshold of Category 2 (b_1), and discrimination singled out Item 29 rather than Item 33 while the threshold of Category 3 (b_2) in Item 15 appeared different from the other two items. Note that the analysis of the misfitting items did not highlight the possibility about the main effects, but the possibility about any interactions is still unknown. More importantly, the findings might suggest that the importance of the separate analysis of the data on the 4-category and the 3-category scales although the responses to Category 1 and Category 2 are the same on both scales.

Table 5.4 Item Parameters and Prompts of the Three Misfitting Items

Items	Difficulty (SE)				Discrimination (SE)
	(b_1)	(b_2)	(b_3)	($b_{overall}$)	
4-category scale					
Q15	-1.35 (0.19)	0.71 (0.15)	1.69 (0.22)	0.35	1.64 (0.23)
Prompt	*Joining a student club on campus is a great way to improve your social skills.*				
Q33	-1.14 (0.15)	0.54 (0.13)	1.53 (0.18)	0.31	1.95 (0.25)
Prompt	*The senior student was talking about his own story of finding an apartment.*				
3-category scale					
Q29	-1.05 (0.14)	0.60 (0.13)	N/A	-0.22	2.19 (0.31)
Prompt	*Regular workouts benefit the body as well as the mind.*				

Related to the prompt characteristics, Item 29 was shorter than the other two, which rather reveals the similarly between Item 15 and Item 33. As a side note, it is interesting to see the much shorter item (Item 29) resulted in similar or higher item difficulty at category level because lengthy EI items are generally expected to be more difficult than shorter items (Miller, 1973; Perkins et al., 1986; Yan et al., 2016). After all, closer examination is needed regarding the rating of grammatical and semantic features that distinguish each category.

In summary, examination of misfitting items disclosed strong potential that fit between observed and model-based scoring is associated with response category levels, and the magnitude of the association varies across the items.

5.3 Discussion and Conclusion

This chapter examined the person and item misfits on the 4- and 3-category EI accuracy scales in relation to person ability and response category levels. First, the results regarding the 22 misfitting examinees revealed 1) fewer misfitting examinees with low proficiency and 2) the greater number of unexpected responses in the lowest and highest categories. In other words, person misfits were caused by medium-level examinees who exactly repeated the prompt of one or more items, or by medium or proficient examinees received the lowest score. Second, the results from the analysis of three misfitting items also indicated 3) associations between item fit and response category levels, negatively with exact repetition in particular, and 4) (some but less clear) associations between item fit and ability although the magnitude of the associations varied across the items.

These findings reasonably raise a question of why. Three possible sources are issues related to rating rubric, item characteristics, and raters' inconsistency. To elaborate, the misfit might have been related to distinguishing grammatical errors and semantic deviations based on quality. In this case, revising the rubric would reduce the discrepancies between model-based and observed scores. Some items might be more vulnerable to misfit because of its characteristics, for example, when items easily elicit examinee mistakes. If that is the case, item revision is recommended. It is also possible that raters were not consistent with other raters or between items when applying the rubric and judging whether the quality of errors or deviation are minor or major.

To illuminate the sources of misfits, more information is needed from qualitative examination of specific errors and deviation observed in examinees' responses as well as from rater feedback (i.e., verbal rating justifications). Responding to the need, Chapter 6 reports the results of qualitative examination on the errors and deviation.

CHAPTER 6. POTENTIAL SOURCES OF MISFITS AND CATEGORY INADEQUACY: INSIGHTS INTO ITEM / SCALE / RUBRIC DEVELOPMENT AND RATER TRAINING (RQ3)

Quantitative examination cannot provide insight into the potential sources of misfit, so responses of misfitting examinees and item misfit were qualitatively analyzed. First, I analyzed the patterns of rating in relation to semantic deviations and grammatical errors in the responses of the misfitting examinees and items (Section 6.1 and 6.2). Next, I examined whether the issues with rating semantic deviations and grammatical errors were found in responses with unexpected scores in non-misfitting items and examinees (Section 6.3). Following the analyses of the sources of the misfits, sources of category inadequacy was examined, focusing on the paraphrase category (Section 6.4). Finally, specific guidelines to minimize the number of misfitting examinees and items and category inadequacy were offered for item development and rater training practices (Section 6.5). The following questions were answered:

RQ3.1 What are the potential sources of the person misfits on the 4- and 3-category EI accuracy scales in relation to rating grammatical errors and semantic deviations? (Section 6.1)

RQ3.2 What are the potential sources of the item misfits on the 4- and 3-category EI accuracy scales in relation to rating grammatical errors and semantic deviations? (Section 6.2)

RQ3.3 Do the issues with rating grammatical errors and semantic deviations observed in misfitting items and examinees exist in item responses with unexpected scores of non-misfitting items or examinees? (Section 6.3)

RQ3.4 What are the potential source of the inadequacy of using the paraphrase category? (Section 6.4)

RQ3.5 What guidelines for item development, scale/rubric revision, and rater training do the qualitative analysis provide to minimize misfits and increase the adequacy of using the paraphrase category?

The qualitative analyses resulted in two main findings. First, the qualitative semantic judgement required by the current scoring method and rubric have likely contributed to unexpected and discrepant scores. Second, grammatical judgement was relatively consistent and caused few issues while some raters considered the quantity of minor errors although this approach is not specified in the rubric. Based on the findings, the testing program has the opportunity to focus on and perhaps more effectively distinguish semantic deviation in item development and rater training, particularly in relation to semantic redundancy and alignment between semantic and grammatical accuracy. Based on the findings, some guidelines for item development, scale/rubric revision, and rater training are offered to minimize misfit and category inadequacy. The findings also suggest the need for exploring alternative scoring methods, which is discussed in detail in Chapter 7.

6.1 Qualitative Analysis: Potential Sources of Person Misfits (RQ3.1)

Table 6.1 indicates that the GRM models run with the scores on the 4-category and 3-category accuracy scales respectively identified 19 and 15 person misfits out of 779 examinees, which resulted in 21 underfits and one overfit in total, of whom 12 examinees misfitted on both scales (See Chapter 4). Among the 252 item responses by the 21 underfitting examinees (i.e., 21 examinees × 12 items) on each scale, 79 responses on the 4-cateogry scale and 48 on the 3-category scales showed discrepancy, which means the observed scores were different from the model-based, expected scores by | ±1.0 | or larger—not by | ±1|. Appendix 6.1 compares observed and model-based item scores of all the responses from the 21 underfitting examinees on the 3-category and 4-category accuracy scales. Eighty-six item responses showed unexpected scores on either scale, including 46 unexpected scores on both scales. Among the 86 responses with unexpected scores, exclusion of 31 responses rated as Category 4 (exact repetition) left 55 item responses, which were qualitatively examined in search of potential sources of the misfit or discrepancy between empirical and model-based scores.

Table 6.2 reports four main possible sources of the unexpected scores: issues with semantic judgement, grammatical judgement, impact of other unexpected scores within the same examinee, and unknown, examinee-related variables. The majority of the responses by misfitting examinees with unexpected scores, 39 out of 55 items, showed issues related to rating semantic deviation while rating grammatical errors posed only a few issues. Approximately 25% of the responses, 14 out of 55, did not demonstrate any issues related with either potential source, eight of which were not flagged when other unexpected scores within the same examinee were adjusted considering the related potential issues, but the other six responses remained unimpacted. Many reasons might account for the misfitting six responses, for example, examinees' attention or interaction with the environment, but not related to rating, rating scales, or rubrics. The next sections present the subcategories of the potential sources.

Table 6.1 Distribution of unexpected scores in misfitting examinees' responses by scale

	Scale		Total	Notes
Frequency	4-category	3-category		
Misfitting examinees	19*	15*	22*	12 misfits * on both scales *one overfit included
Responses with unexpected scores by 21 underfitting examinees	82**	50	86**	46 responses with unexpected scores on both scales; ** 31 unexpected scores rated as Category 4

Table 6.2 Potential sources of responses with unexpected scores by category (N=55)

Response category	Sources of misfit/discrepancy		No rating issue	
	Semantic judgment	Grammatical judgement	Impact of examinee's other unexpected scores	Unknown
3 (paraphrase)	7	0	5	1
2 (minor error/deviation)	2	0	0	1
1 (major error/deviation)	30	2	3	4
Sub-total	39	2	8	6

6.1.1 Sources Related to Semantic Judgement

With closer examination of the 55 highly unexpected responses, Table 6.3 subcategorizes the issues related to semantic judgements into four types: 1) paraphrasing using simpler language, 2) omitting semantically less essential lexis, 3) rating inconsistency about the degree of semantic deviation.

Table 6.3 Types of Potential Sources of Unexpected Scores Related to Semantic Judgement and Selected Examples

No.	Examinee ID	Item No.	Prompt (Changed words in bold)	Response (Changed words in bold)	Rating (Model-based)	GR errors	Words or phrase with issues	Other sources
Main source A. Paraphrasing using simpler language (less complex syntax and more frequent lexis)								
1	F1-40	Q10	Purdue **ranks** second in foreign student **enrollment** among all public schools.	Purdue **has second** in all public universities for **getting** foreign students.	2 (1)	minor	has (vs. rank), getting students (vs. student enrollment)	C2
Main source B. Rating inconsistency regarding omission of semantically less essential lexis								
2	F1-80	Q1	Most students declare their major at the end of their sophomore year **in college.**	Most students declare their major at the end of their sophomore year	3(2)	none	in college	
3	F1-40	Q12	Although **he** did not **review** for the final exam, he scored very high **on that test**.	Although **the student** didn't **study** for the final exam, he scored **really** high.	3 (2)	none	on that test	C1
4	F3-71	Q34	Students can take courses that **have nothing to do with** their major **areas of study**.	Students can take courses that **are not related to** their major.	3 (2)	none	areas of study	A, C1
Main source C1. Rating inconsistency regarding semantic judgement (paraphrase vs. minor deviation)								
5	F4-70	Q37	As you can see on the course schedule, we will not have **a** final exam for this course.	As you can see on the course schedule, we will not have **the** final exam for this course.	3 (2)	none	the final exam (vs. a final exam)	-
6	F4-03	Q44	It’s hard to express your ideas if your language **skills are** low.	It's hard to express your ideas if your language **level is** low.	2 (3*)	none	language level (vs. language skills)	
Main source C2. Rating inconsistency regarding semantic judgement (minor vs. major deviation)								
7	F1-73	Q6	It doesn’t matter if you work alone or in a group **on your homework**.	It doesn’t matter if you work in a group or alone **for this project**.	1 (3)	none	for this project (vs. on your homework)	-

Table 6.3 continued

8	F2-62	Q19	Working part-time will help you develop **time management** skills.	Working part-time will help you develop your **part-time managing** skills.	1 (3)	none	part-time managing skills (vs. time management skills)	D
9	F4-70	Q42	**Students** who enjoy **working** in groups are **more** likely to succeed.	**Student** enjoying **work** in groups are likely to succeed.	1 (2)	minor	likely (vs. more likely)	-
Main source D. Major semantic deviation with no/few errors and high similarity								
10	F1-137	Q8	If you record your lectures, you can **revise** your class notes later.	If you record your lectures, you can **review** your class notes later.	1 (3)	none	review	-
11	F1-137	Q12	Although he did not review for the final exam, he scored very high on that **test**.	Although he did not review for the final exam, he scored very high on that **class**.	1 (3)	none	class (vs. test)	B, C2
12 13	F2-21, F2-72	Q19	Working part-time will help you develop **time** management skills.	Working part-time will help you develop management skills.	1 (2)	none	management (vs. time management)	C2
14	F3-22	Q32	Many students live **off** campus because the rent is much lower.	Many students live **on** campus because rent is much lower.	1 (2)	none	on (vs. off)	-
15	F4-03	Q38	In the event of a **car** accident, you should first stay calm and then call the police.	In the event of a **fire** accident, you should first stay calm and then call the police.	1 (3)	none	car accident (vs. fire accident)	-
16	F4-70	Q48	You should talk to your advisor **if** you are not sure **what** courses to take next semester.	You should talk to your advisor **that** you are not sure **which** courses to take next semester.	1 (3)	none	that (vs. if)	B, C2

Notes: GR=grammar; *flagged on 4-category scale only; **flagged on 3-category scale only

First, rating responses with much lower syntactic complexity and lexical sophistication than the original language of the prompt requires attention (Source A). Example 1 in Table 6.3 (i.e., *Purdue has second in all public universities for getting foreign students*.) shows that the examinee paraphrased *ranks second* and *foreign student enrollment* as *has second* and *getting foreign students*, respectively. This response was rated as 2 (i.e., minor grammatical errors or meaning change). The paraphrased response might convey similar meaning to the original prompt, but the level of simplification is quite substantial. Similarly, but to less degree, Example 4 also shows some difference between the target language used in the prompt (i.e., *have nothing to do with*) and the expression used in the examinee's response (i.e., *are not related to*).

The issue would become more apparent if a lower rating is given to a response that contains major grammatical errors made during an attempt to use the original complex language, *rank* and *enrollment* such as Example 1a below. In addition, when comparing Example 1 with a response rated as the same category (i.e., minor error/deviation) but showed much greater linguistic similarity (e.g., Example 1b), the fairness and construct validity of using the scores and interpretations can be questioned.

<u>Prompt (Item 10)</u>

Purdue ranks second in foreign student enrollment among all public schools.

<u>Responses</u>

Example 1. Purdue **has second** in all public universities for **getting** foreign students. (Category 2. minor error/deviation)

Example 1a. Purdue **was ranked secondly** in foreign student **enrolling** in all public schools. (Category 1, major error/deviation)

Example 1b. Purdue **ranks second** in foreign **student enrollment** among all public **school**. (Category 2, minor error/deviation)

Plus, semantic judgement on *has second* is possibly inconsistent among raters. One similar response with *got second* was rated as 1, which might be *has/got second* can be interpreted related to *having a short period time*, which is a very different meaning.

Another issue arose regarding responses that omitted semantically less essential lexis (Source B). Three examples in Table 6.3 show that the examinees omitted *in college* in Item 1

(Example 2), *on the test* in Item 12 (Example 3), and *areas of study* in Item 34 (Example 4). The omitted word or phrase was semantically less essential because the omitted information can be assumed from the context. For example, in Item 1, *Most students declare their major at the end of their sophomore year in college,* the omission of *in college* led little meaning change because *major* and *sophomore* generally assume of college. Thus, the responses in the source type B were not penalized for semantic deviation and rated as 3 (i.e., errorless/appropriate paraphrase).

However, it is hard to judge whether the omission was due to a failure to fully process the sentence or due to avoidance of redundancy for efficient communication. The gray area could cause an issue for fairness, construct validity, and measurement preciseness because some responses were rated lower than Category 3 because of mistakes made during the repetition of the omitted phrases (e.g., *on collage*, *at the test*, *on test*, *major area of studies*).

Moreover, some raters penalized omitting semantic redundancy when rating the same items of Example 2 to 4 or other items, which raises rating inconsistency and fairness issues within and between items. One rating for Example 4 was Category 2, due to the omission of *areas of study* as indicated in the rater' comment, and similar penalization was found in other responses to the three items. Also, omitting semantically less essential words was treated as minor or major semantic deviation in some other items, as follows:

- omitting *on campus* in *Joining a student club on campus is a great way to improve your social skills* (Item 15)
- omitting *for employers* in *Foreign students are only permitted to work part-time for employers on campus* (Item 26)
- omitting *course* in *When you look at the course schedule, you will see the dates for midterm and final exams* (Item 28)

The inconsistency on whether or not to penalize the omission caused a larger misfit when the expected score was low but not penalized, in other words, when the examinee proficiency was but rated high, or vice versa.

Not surprisingly, most of the rating inconsistency was concerned about judging the degree of semantic deviation. Some inconsistency judging whether the meaning change is little, which is one criterion of appropriate paraphrase or minor (Source Type C1). In Example 5 (Item

37) *the final exam* was rated as appropriate paraphrases of a final exam, but the same replacement was rated as minor deviation/error in other responses to Item 37 or other items. On the other hand, in Example 6 (Item 44), rater judgements varied between paraphrase and minor deviation when comparing *language skills* with *language level* in that only slightly more raters regarded it as minor semantic deviation.

More cases were related to rater judgement on minor versus major semantic deviation (Source Type C2). Example 7, 8, and 9 are the cases rated as major semantic deviations. The boundaries, however, appear to be blurry. In some responses, raters did not regard *project* as major semantic deviation from *homework* of the prompt (Item 6) because *homework* can be often called *a project*, especially when it is a group work. Also, the differentiation of *likely* and *more likely* (Item 42) was considered minor rather than major in other responses because the overall meaning is similar unless the comparison is particularly emphasized. Similarly, it is possible that some raters would find *part-time managing skills* (Item 19) semantically fairly similar to *time management skills* by considering *part-time* as a minor deviation from *time* and *managing skills* as a less frequently used form than *management skills* rather than a major deviation.

While some degree of rating inconsistency is unavoidable in any human rating involving qualitative judgement on linguistic quality, rating inconsistency inherent to the current quality based semantic judgement invites exploration of alternative scales and rubrics, as well.

The final source type related to semantic judgement is responses that featured with major semantic deviation with no/few grammatical errors and high similarity (Source Type D). Example 10 to 16 apply to this source type. In contrast to the three sub-categories of semantic deviation issues (i.e., Source Type A, B, C1/C2), raters consistently assigned a rating of major semantic deviation to the responses in this category because the change or omission of the key or peripheral information changed meaning clearly. For example, *review your class notes* is clearly deviated from *revise your class notes* (Example 10, Item 8), and the same judgement holds valid for other examples: *on campus* versus *off campus* (Example 13, Item 32*), a fire accident* versus *a car accident* (Example 14, Item 38), and *management skills* versus *time management skills* (Example 12, Item 19). In Example 11 (i.e., *on that class* versus *on that test*, Item 12) and Example 15 (i.e., *that you are not sure* versus *if you are not sure*, Item 48), the penalized word conveys peripheral information but created meaning deviation when changed.

Note that, however, the responses in Example 10 to 16 are free of grammatical errors, which is not aligned with the semantic performance. Also, except for the one word that drastically changed the meaning, the responses are linguistically identical (or highly similar) to the prompts. This disparity between semantic performance and overall linguistic similarity, including grammatical accuracy, poses a question of what language proficiency that EI intends to measure is. Whether the construct is implicit knowledge, processing competence, or overall oral proficiency, the one- word difference is a piece of information towards overall meaning that the prompt conveys. When the one-word difference lowers the score from the highest to the lowest (i.e., exact repetition to major semantic deviation), the construct represented by the score is also changed, which resulted in disregarding the language proficiency that enabled the examinee to exactly process and produce the rest of the prompt.

When we compare Example 16 with Example 16a, another response to Item 48, the construct-related issue becomes more prominent. Example 16a was rated higher than Example 16 although it is less similar to the prompt and less accurate because the grammatical errors (i.e., missing *to* after *talk*, *would*, missing *about* before *which*, pluralization, and unnecessary *for*) and omitted part (i.e., *if you are not sure*) were considered as minor. When Example 16a is considered representation of higher proficiency than Example 16, the construct validity of using the score and interpretations can be problematic, or at least, changed.

<u>Prompt (Item 48)</u>

You should talk to your advisor if you are not sure what courses to take next semester.

<u>Responses</u>

Example 16: *You should talk to your advisor that you are not sure which courses to take next semester*. (Category 1. major error/deviation)

Example 16a: *You should talk your advisor which course you would take for next semester*. (Category 2. minor error/deviation)

Rating responses of high linguistic similarity to the prompt as major semantic deviation can also challenges fairness. Although the number of impacted examinees or responses might not be too large, the current rubric possibly penalizes responses and examinees with the discrepancy (e.g., Example 16) than ones without discrepancy (e.g., Example 16a). Example 16a appears less proficient in terms of grammatical accuracy and linguistic complexity than Example 16 but rated higher. The construct validity and fairness issues in these examples are also related to the issue of omitting a semantically less essential lexis. In Example 16, omitting *if you are not sure* was treated as minor deviation because it was peripheral information, but Example 16, which actually reconstructed the part with deviation, was penalized more severely. Similarly, *on the class* in Example 12 was penalized due to the reconstruction from *on the test* but when the information was entirely omitted in Example 3, the omission was treated as appropriate paraphrase. The consequence of the unfair rating can be incremental depending on item difficulty and person ability in that non-alignment more negatively affects the person fit with an easier item and/or an examinee of higher ability.

Reliability is also vulnerable to the adverse effect of the non-alignment. Particularly when the omitted or changed part is not a part of the core structure, whether or not to treat the information as peripheral directs rating, which can be varied among raters.

The complementary alignment between semantic and grammatical accuracy suggests reconsidering the assumptions that the current scoring method and rubric make. The current assumption is that major errors are associated with major semantic deviation. The assumption holds true in many cases because classification of minor and major grammatical errors depends on comprehensibility, in other words, global and local meaning. However, the opposite direction is not necessarily true, as shown in Example 10 to 16.

More fundamentally, it is not always crystal-clear which part can be considered redundant and thus allowed to be omitted without penalization in rating. Key or main information depends on the context. For example, if Example 16 and 16a respond to a question that asks conditions in which one should ask advisor about courses, the information from the *if*-clause should not be omitted. However, EI does not either provide this context or intend to measure examinee’s ability to discern the difference. Rather, EI is designed to measure how accurately and efficiently examinees process and reproduce the given information. The extent to which responses are linguistically similar to the prompt reflects the processing competence. The linguistic similarity cannot be solely measured by one aspect of language proficiency, such as semantic accuracy, but rather via comprehensive consideration of grammatical and lexical reconstruction as well.

The issues with construct validity, fairness, and reliability introduced by the current scoring method and rubric invite to explore alternative approaches to rating semantic deviation.

Considering the issues revolving around the judgement of minor versus minor degree of semantic deviation, semantic judgement from a quantitative perspective might be a good alternative. Specific suggestions will be discussed in Chapter 7.

6.1.2 Sources Related to Grammatical Judgement

The qualitative examination of responses with unexpected scores by 21 underfitting examinees revealed that rating based on grammatical judgement has only a few issues. The analysis led to identification of one type of potential sources of misfit: rating inconsistency regarding the degree of error, minor versus major. Table 6.4 presents the examples of two sub-categories.

Table 6.4 presents two examples that were categorized into inconsistency of grammatical judgement, which was the only responses that had a grammar related issue among the 55 unexpected scores. In Example 17, the grammatical error of using a comparative (i.e., *much more lower*) was rated as a major error. Although there is no hard rule for judging the degree of grammatical error, the current rubric takes how grammatical errors impacts comprehensibility for communication into account. Since the error of using a comparative does not hinder the comprehensibility or global communication, minor error appears to be more appropriate. Although this inconsistency might have been a simple mistake, it seems that rater actually regarded this error as major from the rater justification. It is important to continue to clarify the rating criteria with specific examples during the rater training.

Similarly, the errors of Example 18 were treated as major ones. The raters of Example 18 commented on grammar, but not on semantic deviation, which suggests that the reason for lowering scores to Category 1 is not semantic deviation but grammar errors. However, the grammar errors in Example 18, namely, errors of using plural or agreement (i.e., *students is*, *choose*), tense (*is*, *choose*), and article (before *story*), do not appear to be or were rated as major errors in other responses. This could be a simple mistake, but it is also possible that raters have different viewpoints on which grammar errors interfere with comprehensibility for communication severely.

Interestingly, one of the raters mentioned *lots of errors* as rating justification, which might imply that the quantity of minor errors was possibly the reason why the rater assigned Category 1, major errors, to this response although the quantity of errors is not a criterion for rating in the current rubric. Given that some responses rated as Category 2 have only one error with high linguistic similarity—for example, response with only one error of using an article, agreement, or pluralization—it might not be atypical for raters to subconsciously attend to the quantity as well as the degree. From the fact that the score for Example 18 is discrepant from the model-based score, this example does not support rating responses with numerous minor errors as Category 1 along with responses with major grammar errors. However, considering the wide range of grammatical performance within Category 2, differentiating ratings considering the number of minor errors but still assigning a higher rating than responses with major errors might improve the sensitivity of distinguishing examinees of different proficiency levels. Further discussion is provided in the next section and Chapter 7.

Table 6.4 Types of Potential Sources of unexpected Scores Related to Grammatical Judgement and Selected Examples

No. (Examinee ID)	Item No.	Prompt & Response (Changed words in bold)	Rating (model-based)	GR errors	Words/ phrase with issues
Main source E. Rating inconsistency regarding the degree of errors: minor vs. major					
17 (F3-74)	Q32	Prompt: Many **students** live **off** campus because the rent is much lower. Response: Many **student** live **out of** campus because the rent is much **more** lower.	1 (2)	minor	much more lower (vs. much lower)
18 (F3-72)	Q33	Prompt: The senior **student was** talking about **his own** story **of finding an** apartment. Response: The senior **students is** talking about story **that** he **choose his own** apartment.	1 (2)	minor	is (vs. was)

Note: GR = Grammar

6.2 Qualitative Analysis: Potential Sources of Item Misfits (RQ3.2)

In Chapter 4, the GRM models identified three misfitting items at the 5% significance level: Item 15 and Item 33 on the 4-cateogry scale and Item 29 on the 3-category scale. The responses to the three items were examined to identify the potential issues of rating as well as the current scale and rubric, if any.

6.2.1 Sources Related to Semantic Judgement

The results are presented in Table 6.5. Aligning with the findings from the analysis of misfitting examinees' responses, the responses with discrepancy in the three misfitting items also showed all four types of sources related to rating semantic deviation.

First, simplified paraphrases (Type A) were found in some discrepantly rated responses from the model expectation. In Item 29, *as well as* was paraphrased as *and* (in Example 20a and 20b), which was not penalized in terms of semantic deviation because of marginal meaning change despite the different level of sophistication.

Omitting semantically less essential and/or possibly redundant lexis (Type B) was not penalized, either, for example, omitting *student* in *student club* and *on campus* in Item 15 (in Example 19a, 19b, and 19c) and *student* in *senior student* in Item 29 (in Example 21b). The non-penalized omission introduces concern over fairness because responses that reconstructed the semantically redundant expression is penalized if the reconstruction has an error, as in *in campus* of Example 19d. The non-penalized omission also raises a construct validity issue. Both Example 19a (i.*e., Join a club will improve your social skills*) and Example 19b (i.e., *Joining a student club in campus is a great way to improve your social skills*) were rated as Category 2 due to one minor error but Example 19a is much less sophisticated than Example 19d and much less similar to the prompt. Interestingly, some simplified or omitted expressions were formulaic language, such as *on campus* and *a great way to* in Item 15 and *as well as* in Item 29. The non-penalized omission or simplification degrades the sensitivity to measuring these formulaic languages and interferes with distinguishing examinees of different proficiency levels regarding target formulaic languages.

Table 6.5 Potential Sources of the Three Misfitting Items on 3-Category or 4-Category Accuracy Scales in Relation with Rating, Scales, or Rubric

No. (Examinee ID)	Examinee Response (Changed words in bold)	Rating (Model-based)	GR errors	Words /phrase with issues	Sources of misfit
	Q15. Joining a student club on campus is a great way to improve your social skills.				
19a (F2-11)	Join a **club will** improve your social skills.	2 (1)		club (vs. student club), on campus, will improve (vs. is a great way to improve)	B, C2
19b (F2-3)	Joining a **club** on campus is a great way to improve your social skills.	3 (2)	none	club (vs. student club)	B, C1
19c (F2-150)	Joining a student club is a great way to improve your social skills.	3 (2)	none	on campus	B
19d (F2-68)	Joining a club **in campus** is a great way to improve your social skills.	2 (3)	minor	in campus (vs. on campus)	B, quantity of minor errors
19e (F2-136)	Joining **the** student **clubs** on campus is a great way to improve your **skills**.	1 (3)	minor	skills (vs. social skills)	D
19f (F2-183)	Joining a student club is a **good** way to improve your **English** skills.	1 (3)	none	English (vs. social)	D, C2
	Q29. Regular workouts benefit the body as well as the mind.				
20a (F3-43)	Regular workouts **improve** the body **and** the mind.	3 (2)	none	and (vs. as well as)	A, C1
20b (F3-40)	Workouts **benefits** the body **and** the mind.	2 (1)	minor	and (vs. as well as)	A

Table 6.5 continued

20c (F3-38)	Regular workouts **benefits** the body as well as the mind.	2 (3)	minor	-	quantity of minor errors
20d (F3-33)	Regular **workout benefit to** the body as well as **mind**.	2 (1)	minor	-	quantity of minor errors
20e (F3-126)	Regular workouts **exercise** the body as well as the mind.	1 (3)	none	exercise (vs. benefit)	C2, D
20f (F3-194)	Regular workouts benefit the body as well as the **body**.	1 (2)	none	-	D
20g (F3-98)	Regular **work benefits** the body as well as the mind.	1 (2)	none	work (vs. workout)	D
	Q33. The senior student was talking about his own story of finding an apartment.				
21a (F3_100)	The senior student **is** talking about his own story **about** finding an apartment.	3 (2)	minor	is (vs. was)	rating mistake (grammar)
21b (F3_66)	The **senior** was talking about his own story about finding an apartment.	3 (2)	none	senior (vs. senior student)	B, C1
21c (F3_113)	The senior student **is** talking about his own story of finding an apartment.	2 (3)	minor	-	quantity of minor errors
21d (F3_157)	The senior student was talking about his own story **about** finding **a** apartment.	2 (3)	minor	-	quantity of minor errors
21e (F3-90)	**A** senior **students is** talking about his own story about finding **apartments**.	2 (1)	minor	-	quantity of minor errors

Table 6.5 continued

21f (F3-24)	The senior student **is telling** his own story **about finding his** apartment	2 (2)	minor	-	C2
21g (F3-63)	The senior student was **telling a** story **about** finding an apartment.	1 (2)	none	a (vs. his own)	C2
21h (F3_180)	The senior student was talking about his own story **about owning** an apartment	1 (2)	none	owning (vs. finding)	C2, D
21i (F3_74)	The senior student **is** talking about his **experiment about** finding an apartment.	1 (2)	minor	Experiment (vs. story)	C2, D
21j (F3_30)	The senior student **is** talking about his own story about finding a **campus**.	1 (2)	minor	Campus (vs. apartment)	C2, D

Notes: A – paraphrasing using simpler language; B – omission of semantically less essential lexis; C - rating inconsistency regarding semantic judgement (C1: paraphrase vs. minor deviation; C2: minor vs. major deviation); D – major semantic deviation with no/few errors and high similarity

Rating inconsistency on the degree of semantic deviation (Type C) within the same response was not infrequent in responses with unexpected scores of the misfitting items (e.g., Example 19b, 19f, 20a, 20f, 21b, and 21g). Particularly, comparison of Example 21f versus 21g, two different examinees responses to Item 33 (i.e., The senior student was talking about his own story of finding an apartment), illuminates the inconsistency and corresponding fairness and validity issues. Both responses replaced *talk about* with *tell* and modified determiners (i.e., *his*, *a*), but Example 21g (i.e., *The senior student was telling a story about finding an apartment*) correctly reconstructed the tense of the prompt while Example 21f (i.e., *The senior student is telling his own story about finding his apartment*) did not. However, Example 21f was rated as minor error/deviation while Example 21g, more accurate grammatical reconstruction, was rated as major error/deviation. This inconsistency was likely because the raters viewed "*his own story*" as major deviation from "*a story*", which can be minor to other raters.

Non-alignment between semantic and grammatical accuracy and/or linguistic similarity (Type D) was also found (See Example 19e, 19f, 20e, 20f, 20g, 21h, 22i, and 22j). Most responses with the non-alignment issue were penalized due to major semantic deviation caused by one word. Concerns over construct validity and fairness of scoring arise due to the higher linguistic similarity to the prompt than responses with non-penalized omission or simplification as well as with multiple errors or deviations. These concerns invite reconsidering the criteria for semantic errors.

6.2.2 Sources Related to Grammatical Judgement

In line with the findings from examining misfitting examinees' responses, grammatical judgement did not involve as many issues as semantic judgement. There were a few inconsistencies in rating the degree of grammatical errors in some responses with within the same response scores in the three misfitting items, which was highly likely simple mistakes, such as in Example 21a, where raters did not recognize *is*.

The analysis of responses to the three misfitting items, however, highlighted one interesting point about rating scale of grammatical accuracy. Among the responses rated as minor grammar error, responses with a higher expected score tend to include a smaller number of minor errors than ones with a lower expected score. This might suggest refining the minor error category based on the quantity of minor errors. Specific application is discussed in Chapter 7.

6.3 Qualitative Analysis: Potential Sources of Unexpected Item Scores (of Non-Misfitting Items and Examinees) (RQ3.3)

In order to find out whether the potential sources of misfit found in responses of misfitting examinees and misfits exist in responses with discrepancy in other examinees and items, qualitative examination was conducted on responses with unexpected scores to 45 items, which did not show item misfit (i.e., items except for Item 15, 29, and 33). The analysis revealed similar issues in rating semantic deviation and grammatical errors in the responses with discrepancy. The analysis also revealed some additional sources of rating inconsistency related to grammatical judgement: grammatical redundancy and phonological context. The analysis also found some other sources that are potentially prone to elicit examinees' mistake although not related to rating scales, rubrics, or rating practice. The results are reported in Appendix 6.2. Note that the sources and examples are from the responses with discrepancy between observed and model-based scores. The analysis does not include the responses with little difference between empirical and model-based scores.

6.3.1 Sources Related to Semantic Judgement

The four types of misfit sources related to rating semantic deviation were also found in discrepantly rated responses from the model across the 48 items, for example, non-penalized paraphrasing using less sophisticated language than prompt (in Example 22 and 23), non-penalized omission of less essential lexis (in Example 24a and 24b), rating inconsistency on the degree of semantic deviation (in Example 25a and 25b), and responses with major semantic deviation but high linguistic similarity with their prompt (in Example 26 to 29). Thus, the findings support the issues related to rating semantic deviation and corresponding concerns over fairness and construct validity of score uses and interpretations are not limited to misfitting examinees or items but generally applied to across the test items and examinees to varying degree.

Example 22. <u>*You can*</u> *work alone or in a group on your homework.*

Example 23. *Sometimes, it's good to ask questions in class,* <u>*not*</u> *keeping them to yourself.*

(Responses rated as appropriate paraphrase despite the use of substantially less sophisticated language, by replacing *You can* vs. *It doesn't matter if you* with *You can* (in Item 6) and *as opposed to* with *not* (in Item 25)

Example 24a. *Most students declare their major at the end of their sophomore year*.

Example 24b. *Most students declare their major at the end of their sophomore in college*.

(Responses to Item 1 rated as appropriate paraphrase, which omitted *in college* at the end and *year* in *sophomore year*).

Example 25a. *College students can ride the bus for free as long as they have a valid ID*.

Example 25b. *Graduate students can ride the bus for free as long as they have a valid student ID*.

(Responses to Item 2, which showed inconsistency between paraphrase and minor semantic deviation (regarding *ID* vs. *student ID*), and between minor and major semantic deviation (regarding *graduate students* vs. *collage students*).

Example 26. *Last month, we traveled to Chicago, which is the second largest city in the country*.

Example 27. *First of all, you must attend all the classes to pass this test*.

Example 28. *This university has the third largest campus in the States*.

Example 29. *Most graduate student move out of their dorm after their sophomore years*.

(Responses rated as major semantic deviation due to a one-word difference (*second* vs. *third* in Item 3, *test* vs. *course* in Item 5, *state* vs. *States* in Item 30, and *graduate* vs. *college* in Item 43,))

6.3.2 Sources Related to Grammatical Judgement

Qualitative examination of responses with unexpected scores in non-misfitting items and examinees indicated that rating inconsistency in the degree of grammatical errors, minor versus major errors, was marginal. The analysis also noted a range of grammatical accuracy within the same item score, Category 2, as in misfitting items and examinees. Two responses to Item 43 given below well describe the range within the same category of minor errors in that Example 30a has only one pluralization error while Example 30b has multiple errors.

Example 30a. *Most college students move out of the dorms after their sophomore years*.

Example 30b. *Most the student move out of dorm on the second years.*

(Responses to Item 43 rated as minor errors)

These findings aligned with the results from the analyses of the responses to misfitting examinees and items, which confirmed that concerns and suggestions regarding misfit due to grammatical judgement, if any, can be generally applied to across the items and examinees of the current study.

The examination also identified two additional sources: Grammatical redundancy and phonological contexts. The issue with grammatically redundant language is slightly different from paraphrasing prompts using simpler lexico-grammatical resources (Type A), one of the semantic-judgement related sources because the redundant phrase can be omitted. Example 31 is a response to Item 3 (i.e., *Last month, we traveled to Chicago, which is the third largest city in the country*.). This response omitted *which is* and was assigned to Category 3, appropriate paraphrase, because the omission does not affect the meaning and grammatically omissible. There were some other items that include a grammatically redundant or omissible part, for example, *that* in *The way that English classes are taught here* in Item 9. While it is hard to justify that omitting *that* represents lower grammatical proficiency, the difference lies in the additional grammatical information *which is* includes, agreement and tense. Thus, non-penalized omission of *which is* might raise the same concerns that non-penalized semantically redundant lexis can prompt. With few similar cases in the sample, however, it does not seem to bring pragmatical benefit but worth noting this as another possible type of misfit source.

Example 31. *Last month, we traveled to Chicago, the third largest city in the country*.

On the other hand, it can be beneficial to attend to phonological contexts of items as a source of inconsistency in grammatical judgement and corresponding misfits. Some word combinations were found to be challenging for raters to judge its morphological accuracy or required raters to make extra effort to distinguish whether or not examinees accurately performed bound morphology. The first case was word combination that blended sounds, where the end sound of a word is the same as the first sound of the following word, as in *ranks second* in Item 10 (Example 32). Because the two /s/ sounds are supposed to be pronounced as one sound, it is hard to distinguish whether an examinee says *rank second* or *ranks second*. This distinction is important because *ranks second* receives the highest rating, Category 4, but rank second is rated as Category 2, when the rest is also exactly repeated. This problem can be easily avoided with simple modification, for example, *second* to *first*.

Example 32. *Purdue rank**s s**econd in foreign student enrollment among all public schools*. (Item 10)

Another case is to rate morphological grammatical accuracy in bigrams that include consonant clusters when two words are used together. The four examples given below (Example 33 to 36) show that *traveled*, *borrowed*, *friends*, and *students*, the morphological aspect of which needs rating, become harder to distinguish when its end sound is linked to the first sound of the following word, which results in a consonant cluster.

Example 33. *Last month we travele**d t**o Chicago, which is the third largest city in the country*. (Item 3)

Example 34. *Borrowe**d b**ooks from the library must be returned or renewed by the posted due dates.* (Item 23)

Example 35. *Meeting people and making friend**s sh**ould be an important part of your college life.* (Item 46)

Example 36. *Student**s sh**ould know that they can also borrow books from libraries of other schools.* (Item 47)

Because consonant clusters are a natural part of English language and are often paid central attention in L2 English pronunciation and prosody instructions, it is ideal to include words with consonant clusters in EI prompts. However, when consonant clusters result from bigram and create difficulty of rating for bound morphologies, raters should pay attention to rating these bigrams to minimize rating inconsistency. While rater training can raise raters' awareness of the issue related to rating consonant clusters, simple revision to remove this type of consonant cluster is also an option because practicality, efficiency, and reliability are important benefits of using EI, to which short and easy rating is key.

6.3.3 Sources of Examinee Mistakes Related to Prompt Characteristics

Additionally, the analysis indicated that some other issues related to prompt characteristics might have caused examinees to make mistakes rather than errors.. First, it appeared that examinees were more susceptible to make information about numeric information and time, for example, Item 3 (i.e., *Last month, we traveled to Chicago, which is the third largest city in the country*) and Item 3 Item 30 (i.e., *This university has the third largest campus in the state*). Meaning differences related to these issues lowered the item scores although the grammatical structure and accuracy was not degraded. While rating or rating criteria itself do not have any issues, if the incorrect information is simple cognitive mistake rather than language proficiency, it can negatively affect the construct validity of the score uses and interpretations. Particularly, the incorrect numeric information in the responses were treated as major semantic deviation in some unexpected scoring.

Words with similar sounds in a paralleled structure in a prompt seems to have caused examinees to make simple mistakes. For instance, some responses with unexpected item scores showed switched use of *borrow* and *buy* in Item 4 (i.e., *By the way, you can always borrow textbooks from the library or buy them online*) and *returned* or *renewed* in Item 23 (i.e., *Borrowed books from the library must be returned or renewed by the posted due dates*) In addition, a small mistake in differentiating subtle pronunciation was found to be able to lower an item score from full score (i.e., exact repetition) to Category 1 (i.e., minor deviation) such as the sound of one bound morphological item, /ts/ when *state* was pronounced as *States* in Item 30 (i.e., This university has the third largest campus in the state). This can affect the construct that the test is intended to measure from global to very local proficiency by focusing on the differentiation of local pronunciation.

A similar concern was found with words with similar meaning used in the same prompt, for instance, *class* and *course* in Item 5 (i.e., *First of all, you must attend all the classes to pass this course*) and *look* at and *see* in Item 21 (i.e., *You can look at the course schedule to see the dates for midterm and final exams*). While fine-graded differentiation of these similar words is important, if the differentiation is not the main knowledge intended to be tested in EI, examinees who made a mistake of switching the two similar words in their responses can cause the change in the construct measured in the test.

Examinees also seem to have been more vulnerable to make mistakes when dealing with long word combination such as *time management skills* in Item 19 (i.e., *Working part-time will help you develop time management skills*) While it is possible that examinees do not know this multi-word, this type of multi-word that consists of a series of nouns might be susceptible to tap into a different cognitive aspect, memory, as what is required to remember a shopping list. This was particularly problematic because missing *time* in *time management skills* was rated as major deviation.

Finally, including a proper noun potentially introduces an issue because of the possibility of activating real-world knowledge instead of using language proficiency when responding to a prompt. For example, in Item 3 (i.e., *Last month we traveled to Chicago, which is the third largest city in the country*), an examinee might have been interrupted by their pre-existing real-world knowledge about *Chicago* when they were confused about *third* versus *second* or *country* versus *the US*. The former case was rated as major deviation and the latter case as paraphrase, which is still lower than exact repetition.

6.4 Potential Sources of Category Inadequacy (RQ 3.4)

Qualitative examination of responses with unexpected scores also provided insight into understanding adequacy of using the paraphrase category. As discussed in Chapter 4, one issue with the 4-category accuracy scale was that the paraphrase category did not provide substantial unique information, except for four items (Item 5, 8, 9, and 14). The information from the paraphrase category considerably overlapped with information from the minor error/deviation category or exact repetition. In other words, in most items, the paraphrase category was not useful for distinguishing examines of different proficiency levels because proficiency represented by paraphrase overlapped with proficiency indicated by exact repetition or minor error/deviation. Qualitative analysis suggests that the lack of distinguishability can be attributed to rating criteria for the paraphrase and exact repetition categories. The issues with rating criteria are presented with respect to two criteria, first regarding the distinguishing the paraphrase form minor error categories, and second, from exact repetition.

6.4.1 Distinguishability of Proficiency Between Appropriate Paraphrase and Minor Error/Deviation

The previous sections in this chapter pointed out non-penalized simplification and omission due to little semantic changes can cause unexpected and discrepant scores and person and item misfits. The same issues can be the sources of inadequacy of using the paraphrase category. The influence can be more intensive when the original language is sophisticated and thus the item is difficult, as in Example 23 (i.e., *Sometimes, it's good to ask questions in class, not keeping them to yourself*) in Section 6.3, a response to Item 25 rated as appropriate paraphrase. This response changed *as opposed to*, which is the main contributor to the high item difficulty, to *not*. Additional examples (Example 37a, 38a, and 39a) presented below also elucidate the gaps of linguistic sophistication between the original expressions and simplified paraphrases. Since linguistic sophistication is a key to L2 proficiency, when the paraphrase category includes less sophisticated responses, proficiency represented by the paraphrase category is less distinguished from its lower adjacent level, the minor error/deviation category. The decreased distinguishability would reduce the unique information from the paraphrase category and increase the amount of overlaps in the information between both categories.

Example 37a. *It's wonderful that English teachers know their students quite well*. (Rated as 3, appropriate paraphrase)

Example 37b. *The wonderful thing about English teachers are that they know their students quite well*. (Rated as 2, minor error)

Example 37c. *Wonderful thing about English teachers is that they know their students quite well*. (Rated as 2, minor error)

(Responses to Item 16 (i.e., The wonderful thing about English teachers is that they know their students quite well.))

Example 38a. *You can tell me your questions about the final project during my office hours.* (Rated as 3, appropriate paraphrase)

Example 38b. *You can tell me what questions do you have about the final project during my office hours.* (Rated as 2, minor error)

(Responses to Item 22 (i.e., You can tell me what questions you have on the final project during my office hours))

Example 39a. *Foreign students can only work part-time for employers on campus*. (Rated as 3, appropriate paraphrase)

Example 39b. *Foreign students are only permitted for work part-time for employers on campus*. (Rated as 2, minor error)

(Responses to Item 26 (i.e., Foreign students are only permitted to work part-time for employers on campus.)

When the non-penalized, simplified responses are compared with non-simplified but inaccurate repetitions, the proficiency representation of the paraphrase category is more problematic, which brings the discussion of accuracy versus complexity. In Example 37b, 37c, 38b, and 39b, the reconstructed responses demonstrated the same level of syntactic complexity and lexical sophistication as of the prompts but include one minor error. Thus, non-penalized simplification led to overly prioritizing accuracy over complexity and misrepresenting the construct, L2 proficiency. This misrepresentation lowered the distinguishability of the paraphrase category.

In addition, the errors made in the course of reconstructing the prompt in the four non-simplified examples cannot be measured in the simplified example, namely, examinees' agreement with a third person singular subject (Example 37b), article use (Example 37c), word order in an indirect question (Example 38b), and preposition use (Example 39b). These grammatical errors or mistakes often distinguish examinees of different proficiency in the current sample. Thus, accuracy rating in the non-penalized simplified responses is not precise, which also degrades the distinguishability of the category.

Non-penalized, semantically less essential lexis in responses rated as appropriate paraphrase also negatively affects the adequacy of using the paraphrase category in a similar way. Because the omitted language allows responses, otherwise rated as minor error/deviation, to be assigned as appropriate paraphrase, the proficiency levels endorsed by the paraphrase category overlaps more with those by the minor error/deviation category. In conclusion, to increase the adequacy of using the paraphrase category, it is recommended to revise the current rubric and rating practice and consider complexity as well as grammatical and semantic accuracy for the paraphrase category.

6.4.2 Distinguishability of Proficiency Between Appropriate Paraphrase and Exact Repetition

The lack of unique contribution of the paraphrase category can also be attributed to the overlapping information between paraphrase and exact repetition. That is, the proficiency represented by appropriate paraphrase is not clearly distinguished from the proficiency indicated by exact repetition. Table 6.6 presents the three possible types that the qualitative analysis of responses rated as paraphrase identified: 1) switching the order of parallel items, 2) omitting or adding redundant grammatical items, and 3) using interchangeable grammar. These three types of grammatical paraphrase make no change, or marginal at best (Type C), in either semantic accuracy or grammatical complexity.

Table 6.6 Frequent grammatical changes in the paraphrase category

No.	Item No.	Prompt and Source Type	change in responses of the paraphrase category
		A. Switching the order of parallel items	
40	6	*It doesn't matter if you work alone or in a group on your homework.*	- *in a group or work alone*
41	23	*Borrowed books from the library must be returned or renewed by the posted due dates.*	- *renewed or returned*
		B1.Omitting grammatically redundant items	
42	9	*The way that English classes are taught here might differ from the way in your country.*	- *that* (omitted relative adverb)
43	27	*Students can keep the books that they borrow from the library for a semester.*	- *that* (omitted relative pronoun)
44	14	*Before you arrive on campus, you need to make sure that you have a place to live.*	- *that* (omitted objective complementizer of a verbal phrase)
45	47	*Students should know that they can also borrow books from libraries of other schools.*	- *that* (omitted objective complementizer of a verb)
		B2. Adding omittable grammatical items	
46	19	*Working part-time will help you develop time management skills.*	- *to develop* (reconstructed omitted *to* of a *to* infinitive)

Table 6.6 continued

47	9	*The way that English classes are taught here might differ from the way in your country.*	- *the way they are taught* (reconstructed the omitted part due to repetition in a paralleled structure)
		C. Using interchangeable grammatical options in the context	
48	42	*Students who enjoy working in groups are more likely to succeed.*	- in a group (interchangeable singular and plural nouns)
49	43	*Most college students move out of the dorms after their sophomore year.*	- their dorms (interchangeable possessive and definite article)
50	44	*When you take courses here, attendance often counts as a part of the final grades.*	- part of (interchangeable countable and non-countable noun)

The first type of paraphrase (Type A) was to switch the order of the two parallel items (Example 41 and 42). In Example 41, for instance, instead of *returned or renewed* of the original prompt, the examinee responded as *renewed or returned*. The switch lowered the score from exact repetition to paraphrase, but it is hard to justify why the paraphrase is a representation of lower proficiency. If it is not, the proficiency levels represented by the two categories overlap and the unique information from the paraphrase category is reduced.

The second type of paraphrase (Type B) was concerned about grammatically redundant items. Examinees sometimes did not repeat omissible grammatical items, such as -*that* as a relative adverb (in Example 42), relative pronoun (in Example 43), and complementizer (in Example 44 and 45). Lowering item score for these paraphrases can reduce the distinguishability of the items. On the other hand, some examinees reconstructed parts that were omitted in the prompt because the items were redundant (e.g., *to* infinitive in Example 46) or repeated in a parallel structure (e.g., a repeated clause in Example 47). The reconstructed items suggest two important points. First, repetition in EI reflects reconstructed language based on examinee's language proficiency rather than mechanical verbatim. If the grammar or knowledge on the omitted structure had not existed and been processed the examinees would have reconstructed the omitted part. Second, the grammatical reconstruction questions whether or not the language proficiency level represented by this type of paraphrase is lower than the proficiency level indicted by exact repetition. Because examinees are highly likely to have possessed, accessed, and processed the knowledge that the given EI item tested, it is reasonable to assume that the responses rated as paraphrase indicated as high proficiency as exactly repeated responses did. This overlap causes reduced unique information from the paraphrase category.

The final category is the use of interchangeable grammatical options in the given context. This type is different from Type A and Type B because the interchangeability depends on the context of the sentence. The compared grammatical points in Example 48 to 50 are singular versus plural noun (in Example 48), a possessive versus a definite article (in Example 49), and countable versus noncountable noun (in Example 50). These grammatical changes are not always considered interchangeable. However, in the given items, it might be hard to support the differentiation of proficiency levels between one and the other in the three grammatical pairs, or it is very marginal at best.

In summary, the qualitative examination of the responses rated as paraphrases noted three issues in the rating criteria that might lower the distinguishability of the paraphrase category by contributing to overlap in the proficiency levels represented by the paraphrase category and its adjacent categories. Implications for rating scale, rubrics, and rater training are provided in the next section.

6.5 Guidelines for Developing/Revising Items, Scales, and Rubrics and Training Raters (RQ3.5)

This chapter reported the results from the qualitative analysis in search of the sources of unexpected and/or discrepant scores in the responses in the 4-cateogry and 3-cateogry EI accuracy scale. Regardless of their misfit status, discrepantly scored responses compared with the model-based, expected values, although not all the responses, showed issues related to semantic and grammatical judgement in regard to either rating criteria or inconsistency of applying the criteria. Based on the findings from the qualitative analysis of misfit sources, some practical guidelines were proposed for item development, scale/rubric revision, and rater training to minimize unexpected scores and misfits and increase the adequacy of the category usage, particularly the paraphrase category.

The current testing program can consider the following guidelines and suggestions for revising rubric without changing items or scales.

- Guidelines and suggestions for revising (and developing) rating criteria of the rubric

1. Penalize paraphrases using less sophisticated language in complexity and lexis even if meaning change is little rather than rating them as appropriate paraphrase.
2. Penalize omitting semantically redundant information or less essential even if the information can be assumed when omitted.
3. Adjust the degree of penalization considering overall linguistic similarity when only small portion of the prompt (i.e., a part of the word, one or two words) was changed or omitted rather than rating the responses as major deviation.
4. Give an equal rating to words or phrases whose order is switched in a paralleled structure if the order shift does not lead to meaning change and grammatical inaccuracy as the ones whose order is not switched.
5. Give an equal rating to grammatically correct responses that omit grammatically redundant items as the ones that do not if differences in meaning and complexity are little.
6. Give an equal rating to grammatically correct responses that reconstruct grammatically redundant items as ones that do not reconstruct if differences in meaning is slight.
7. Give an equal rating to responses that replace an item in a prompt with grammatical items interchangeably used in the context of the given prompt with little meaning change as the ones that do not.

In the meantime, the guidelines and suggestions are presented for rater training and rating practicing, as follows:

- Guidelines and suggestions for rater training and rating practice:

1. For higher inter-rater and intra-rater consistency,
   - provide a list of omitted and changed grammatical and semantic items in responses with ratings for raters
   - offer a list of grammatical or semantic points that raters frequently make a mistake about for raters (e.g., bound morphologies, articles)
   - check rater justifications for the decision on the degree of semantic deviation
2. For higher rating consistency within items,

- assign rating by items rather than by examinees, which is expected to help lower rater's cognitive load when ratings
- analyze rater performance on item-level in addition to on test-level for rater training

3. For higher rating consistency between items,
   - make sure the same criteria are applied to the same way across items
   - analyze item statistics for rater training

For future item revision and development, the followings are recommended for the current and other testing programs.

- Guidelines and suggestions for item revision and development

Avoid using the following:

1. a word or phrase that can be omissible due to semantic redundancy, such as the ones that can be assumed by lexical items in the prompt
2. omissible grammatical items that can lower the syntactic complexity when omitted, or that frequently elicit grammatical errors
3. parallel items whose order is not clear
4. numeric number or time information that can bring major meaning deviation when changed
5. words with similar sounds in a paralleled structure
6. words that can lead to major meaning difference with subtle difference of pronunciation
7. words with similar meaning in the same prompt
8. multi words that consist of three or more words of the same part of speech
9. pronouns that can activate real-world knowledge
10. a word of which morphological features are hard to be differentiated because the end sound of the word is the same as the first sound of the following word

Finally, the current and other testing programs can consider the followings for future revision and development of scales and scoring methods.

- Guidelines and suggestions for scale/scoring methods revision and development:
  1. Combine exact repetition and paraphrase particularly when:
     - the independent use of the two categories does not serve the purposes of the test better than the combined use of the categories
     - training raters to include exceptions for exact repetition (i.e., rubric guideline 4 through 7 above)
  2. Divide the minor error category into two based on the quantity of the errors
  3. Categorize the semantic deviation based on the quantity rather than the quality

These guidelines are some recommendations to minimize unexpected/discrepant scores and item/person misfits and improve the adequacy of category usage, rather than required steps or exhaustive lists.

6.6 Discussion and Conclusion

Chapter 6 presented the results of the qualitative examinations in search of sources of unexpected scoring and misfitting examinees/items. Overall, issues related to the rater judgement on the degree of semantic deviation seemed to contribute more than grammatical aspects. Treating omission, simplified language, and non-alignment between semantic deviation and linguistic similarity were suggested as main aspects to consider in order to minimize unexpected scores. Some of these aspects and others were found to have potentially lowered the adequacy of the use of the paraphrase category independently from the exact repetition. Based on the findings, guidelines and suggestions were posed for developing and revising rubric, scales, and items as well as practicing rating and rater training. Although the findings and guidelines were constructed based on the data from one local test, implications can be made for testing program in a similar context and population, as well as testing and rating beyond the EI test, for example, sentence-level assessment in ESL and EFL settings. The next chapter elaborates the specific applications of the proposed scale revision.

CHAPTER 7. REVISED SCALES AND SCORING METHODS

Chapter 6 suggested considering quantity-based approaches to revising scales and scoring methods because unexpected scores and misfits on the 4-category and 3-category accuracy scales can be attributed to the rating criteria of semantic judgement based on the degree or quality (i.e., minor versus major). Also, the range of variations in the responses rated as minor error was substantially wide. This chapter introduces the revised scoring methods/rubrics for the original 3- and 4-category EI accuracy scales (Section 7.1) and reports the results of IRT model performances on the original and alternative accuracy scales and rubrics (Section 7.2 and 7.3). The following questions were answered:

RQ4.1 What are the alternative EI accuracy scales and rubrics that address the issues of rating criteria in semantic and grammatical judgement? (Section 7.1)

RQ4.2 Do the quantity-based, alternative EI ordinal scales and rubrics perform better than the quality-based, original EI ordinal scales and rubrics? (Section 7.2)

RQ4.3 Does the best fitting ordinal scale perform better than binary scale? (Section 7.3)

By applying a frequency-based approach, six ordinal scales were proposed, as well as one binary scale, which rates whether semantic or grammatical deviation or error exists. The comparison of performance among the nine scales and rubrics suggested that the best-performing model was the scale/rubric that combined appropriate paraphrase and exact repetition into one category and employed the frequency-based approach to judging both semantic and grammatical accuracy. Interestingly, the binary scale was also fit well in general, outperforming some ordinal scales, although not as well as the best-performing ordinal model. However, the binary scale was found to be least precise among all the scoring options.

7.1 Scale and Rubric Revision Using Quantity-Based Scoring Methods (RQ 4.1)

The qualitative analysis of the responses with unexpected/discrepant scores implied that measuring accuracy based on the degree of semantic deviation in EI potentially causes unexpected scoring and misfits in persons and items. To address the issue, I proposed alternative rubrics and scoring methods that approach semantic deviation and minor errors quantitatively rather than qualitatively. The alternative scoring methods were applied to the 4-category and one 3-category scales, which resulted in six different modification options.

7.1.1 Scoring Based on Frequency of Semantic Deviation

Table 7.1 provides the alternative rubrics based on the rating criteria that judges the frequency of semantic deviation on the 4-cateogry and 3-category scales—henceforth, frequency-based semantic deviation (FSD) scoring, rubrics, or scales. Instead of classifying semantically deviated responses into minor versus major degree of semantic deviation, the alternative method considered the quantity of semantic deviation, in other words, the number of the semantic deviation points. The FSD rubrics assign Category 2 (i.e., minor semantic deviation in the current rubric) to responses that have four or fewer replaced (rather than paraphrased), added, or omitted words, while responses with five or more semantically deviated words are rated as Category 1 (i.e., major semantic deviation in the current rubric). Note that in FSD scoring, semantic deviation is judged based on lemmas, and thus bound-morphological variations of the same lemma are not penalized for semantic deviation, but for grammatical deviation or errors. For example, when responding to Item 4 (i.e., *By the way, you can always borrow textbooks from the library or buy them online*), using *textbook* instead of *textbooks* is not penalized for semantic accuracy but for grammatical accuracy only. Adding re- or suffixes with an opposite meaning, such *un-*, *im-*, or *-less*, is an exception, and should be treated as semantically deviated words, but such cases were not found in the current data.

The criteria of four or fewer semantically deviated words versus more than four words in FS scoring was selected considering that the average number of the semantically deviated words was four among the responses to the two items with the median item difficulty of the 48 items that were rated as Category 1 and Category 2. When sentence length is word-based, that is, length measured based on the number of words, four words accounted for 40% of the shortest items, such as Item 29 (i.e., *Regular workouts benefit the body as well as the mind*) and Item 30 (i.e., *This university has the third largest campus in the state*), and approximately 25% of the longest prompts, which is Item 28 (i.e., *When you look at the course schedule, you will see the dates for midterm and final exams*).

Table 7.1 The Use of FSD Scoring in rubrics on the 4-category and 3-category EI accuracy scales

Category	Current scoring	FSD Scoring
4*	exact repetition	Category 4 of the current scoring *
3	appropriate/errorless paraphrase	Category 3 of the current scoring
2	(a) minor grammatical or (b) semantic deviation (or errors)	(a) minor grammatical errors or (b) *one to four* semantic deviation points**
1	(a) major grammatical or (b) semantic deviation (or errors), or (c) irrelevant / no / incomprehensible responses	(a) major grammatical errors, (b) *more than four* semantic deviation points**, or (c) no / incomprehensible responses

Notes: * Category 4 does not exist on the 3-category scale because the category is combined with Category 3; **semantic deviation points - replaced, omitted, or added words compared to the given prompt rather than paraphrasing the prompt

7.1.2 Scoring Based on Frequency of Grammatical Error / Deviation

Similarly, a frequency-based approach was used to address the potential issue of the wide variations of grammar errors and deviation in the minor error category. The original scale and rubric did not specify grammatical deviation compared with grammatical errors, which was not needed because the differentiation does not affect the scoring decision. However, in this chapter, I differentiated the two criteria. Grammatical errors are incorrectly used grammar while grammatical deviation is correctly used grammar but different from the given prompts. For example, when responding to Item 33 (i.e., *The senior student <u>was</u> talking about his own story of finding an apartment*), when an examinee missed *was*, the difference was considered as a grammatical error, but replacing *was* with *is* was treated as grammatical deviation. Although it should be acknowledged that such and other grammatical changes accompany meaning changes, when the lemma was identical, the change was viewed as grammatical.

Table 7.2 shows scoring of minor grammatical error and deviation based on the frequency in the rubrics on the 4-category and 3-cagegory EI accuracy scales, which are henceforth referred to as frequency-based grammar (Error or) Deviation (FGD) scoring, rubrics, and scales. The FGD scoring method further classifies the minor error category of the current rubric into responses with one or two minor grammatical error/deviation and responses with more than two. Note that the use of the FGD scoring method adds one more category to the current 4-category and 3-category rubrics and scales.

Table 7.2 The Use of FGD Scoring in rubrics on the 4-category and 3-category EI accuracy scales

Category	Current scoring	FGD Scoring
4*	exact repetition	Category 4 of the current scoring *
3	appropriate/errorless paraphrase	Category 3 of the current scoring
2	(a) minor grammatical or (b) semantic deviation (or errors)	(a) *one or two* minor grammatical errors / deviations or (b) minor semantic deviation
1	(a) major grammatical or (b) semantic deviation (or errors), or (c) irrelevant / no / incomprehensible responses	(a) *more than two* minor grammatical errors / deviations or (b) minor semantic deviation
0	None	Category 1 of the current scoring

Note: * Category 4 does not exist on the 3-category scale but is collapsed into Category 3

7.1.3 Scoring Based on Frequency of Semantic and Grammatical Error / Deviation

The third option of alternative scoring methods applied the frequency-based approach to scoring both semantic and grammatical error / deviation to rubrics on 4-category and 3-cateogry scales, which are henceforth referred to as frequency-based semantic and grammar deviation (and error) (FSGD) scoring, rubrics, and scales. The incorporation of both modifications into one rubric was proposed for more precise recommendation because even if both FSD and FGD are found to perform better than the current rubrics and scales individually, the combined use might not result in better performance. On the other hand, it is possible that FSGD can perform better but the independent use of the two scoring methods might not bring benefits. Table 7.3 presents the descriptors of the FSGD rubrics and scales. Note that the use of the FGD scoring method adds one more category to the current 4-category and 3-category rubrics and scales due to the sub-categorization of the minor grammatical error or deviation category.

Table 7.3 The Use of FSGD Scoring in rubrics on the 4-category and 3-category EI accuracy scales

Category	Current scoring	FGD Scoring
4*	exact repetition	Category 4 of the current scoring *
3	appropriate/errorless paraphrase	Category 3 of the current scoring
2	(a) minor grammatical or (b) semantic deviation (or errors)	(a) *one or two* minor grammatical errors / deviations or (b) *one to four* semantic deviations
1	(a) major grammatical or (b) semantic deviation (or errors), or (c) irrelevant / no / incomprehensible responses	(a) *more than two* minor grammatical errors / deviations or (b) *more than four* semantic deviations
0	None	Category 1 of the current scoring

Note: * Category 4 does not exist on the 3-category scale but is collapsed into Category 3

7.1.4 Binary Scale

The final scoring method is binary scale. A binary scale was included as an option for four reasons. First, a binary scale has a pragmatic advantage because of its simplicity. Second, binary scale is a one way of avoiding the potential issues of inconsistency and preciseness of scoring semantic and grammatical deviation and error, as well as related validity issues. In terms of exemption of course enrollment, one of the purposes of the current test, distinguishing advanced and intermediate levels of L2 English performance is most important. Finally, literature (e.g., Yan et al., 2016) found ordinal scoring performs better than binary scoring but evidence from direct comparison is little. The binary scale for this study was modification of the current 4-category and 3-category scales. The scale combines exact repetition and appropriate paraphrase in to one category, and minor and major deviation and errors into the other. Table 7.4 shows how the binary scale operates to measure semantic and grammatical accuracy.

Table 7.4 Scoring Rubric on the Binary EI Accuracy Scale

Category	Current Ordinal Scales	Binary Scale
4*	exact repetition	none
3	appropriate/errorless paraphrase	none
2	(a) minor grammatical or (b) semantic deviation (or errors)	(a) exact repetition or (b) appropriate/errorless paraphrase
1	(a) major grammatical or (b) semantic deviation (or errors), or (c) irrelevant / no / incomprehensible responses	(a) minor grammatical or (b) semantic deviations (or errors), (c) major grammatical or (d) semantic deviation (or errors), (e) irrelevant / no / incomprehensible responses

Note: * Category 4 does not exist on the 3-category scale but is collapsed into Category 3

7.2 Comparison of the Revised Rubrics and Scales: Ordinal Scales (RQ 4.2)

This section reports the performance of the IRT models that were respectively conducted on scores on the eight ordinal scales. Table 7.5 shows the compared scales: two scales based on deviation/error quality and six scales based on deviation/error frequency. The performance of the eight scoring methods was evaluated by multiple indices at test-, item-, and examinee level, that is, model fit, item parameters, item/person misfit, marginal reliability, and preciseness.

Table 7.5 Eight Ordinal Scales/Rubrics/Scoring Methods Compared for the Study

Scale/Rubric	Number of category	Use of the paraphrase category	Rating Criteria of Deviation / Error	
			semantic	grammatical
4C (current rubric)	4	independent	quality-based	quality-based
4C FSD	4	independent	frequency-based	quality-based
5C FGD	5	independent	quality-based	frequency-based
5C FSGD	5	independent	frequency-based	frequency-based
3C (current rubric)	3	collapsed	quality-based	quality-based
3C FSD	3	collapsed	frequency-based	quality-based
4C FGD	4	collapsed	quality-based	frequency-based
4C FSGD	4	collapsed	frequency-based	frequency-based

Notes: 5C : 5-category; 4C : 4-category; 3C : 3-category

7.2.1 Best-fitting Ordinal Model

The results from the examination of test-level indices are presented in Table 7.6. Table 7.6 shows the best and least fitting model performance of each of the eight scales/rubrics. The findings indicated that eight ordinal scales/rubrics fit well across the four forms in general, in that RMSEA values were 0.07 or lower, TLI and CFI were 0.96 or higher, and the p-values were non-significant. There were only two cases that can be considered as an acceptable fit rather than a good fit, which are the 5-category FSGD rubric on Form 4 (RMSEA = 0.08, TLI = 0.85, CFI = 0.91, *p* = 0.06) and the 4-catogory FGD rubric on Form 3 (RMSEA = 0.09, TLI = 0.92, CFI = 0.95, *p* = 0.01). These two cases included low RMSEA values and/or a significant p-value. SRMR values did not vary considerably across the rubrics and forms.

Table 7.6 Model Fit of the Eight Ordinal Scales/Rubrics (Best and Worst Values Across the Four Forms), (N=360 = 90 × 4 forms)

Scale/Rubric	Paraphrase as a combined category				Paraphrase as an independent category			
	3C	3C-FSD	4C-FGD	4C-FSGD	4C	4C-FSD	5C-FGD	5C-FSGD
Best fit								
Form	3	1	1	1	4	1	1	2
M^2	38.97	36.12	20.88	13.94	20.72	24.00	17.16	12.13
LLV	-882.21	-778.57	-1123.26	-1082.16	-1110.19	-971.47	-1313.33	-1259.16
Parameters (*df*)	36 (42)	36 (42)	48 (30)	48 (30)	48 (30)	47 (31)	59 (19)	60 (18)
p (M^2)	0.60	0.73	0.89	0.99	0.90	0.81	0.58	0.84
RMSEA (M^2)	0	0	0	0	0	0	0	0
SRMR (M^2)	0.06	0.06	0.06	0.05	0.07	0.07	0.06	0.05
TLI	1.01	1.02	1.15	1.22	1.06	1.05	1.04	1.14
CFI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Least fit								
Form	4	4	3	4	1	4	4	4
M^2	49.90	59.80	51.03	37.39	32.59	37.70	18.93	28.18
LLV	-1209.32	-776.42	-1131.51	-1115.93	-1092.79	-946.07	-1380.36	-1283.64
Parameters (*df*)	36 (42)	36 (42)	48 (30)	48 (30)	47 (31)	48 (30)	60 (18)	60 (18)
p (M^2)	0.19	0.04	0.01	0.17	0.39	0.16	0.40	0.06
RMSEA (M^2)	0.05	0.07	0.09	0.05	0.02	0.05	0.02	0.08
SRMR (M^2)	0.07	0.07	0.07	0.06	0.07	0.07	0.07	0.06
TLI	0.98	0.96	0.92	0.96	0.99	0.96	0.99	0.85
CFI	0.98	0.97	0.95	0.97	0.99	0.97	0.99	0.91

Notes: LLV: log-likelihood value

Although the model fit indices showed that overall, the eight scales/rubrics are comparable with only two exceptions, the comparison of information criterion values across the scales/rubrics revealed that modifying semantic criteria improved the model fit. Table 7.7 presents the AIC and BIC values of the eight scales/rubrics across the four forms. When comparing the original quality-based semantic scoring to its corresponding frequency-based method, that is, 3C versus 3C-FSD and 4C versus 4C-FSD, across all four forms, the AIC and BIC values noticeably decreased. The minimum decrease was 214.45 in AIC and 216.95 in BIC in Form 3 between the 4-category original and semantically frequency-based scales (i.e., 4C vs. 4C-FSD), and the maximum values, 889.8 in AIC and 919.8 in BIC, were found in Form 4 between the 3-category original and alternative scales (i.e., 4C vs. 4C-FSD). The decrease is very substantial because a decrease of 10 in BIC is considered strong evidence (Raftery, 1995).

On the other hand, refining minor grammatical errors increased the AIC and BIC values, which is expected because the added category per item results in 12 more parameters. Information criterion statistics, particularly, BIC, penalize a more complex model. However, the AIC and BIC values were decreased from the models of the scales/rubrics based on frequency-based grammatical judgement (i.e., 3C-FGD, 4C-FGD) to the models of alternative scales with a frequency-based approach to rating both grammar and semantic deviation and errors (i.e, 3C-FSGD, 4C-FSGD). The decreases were smaller than those occurring between original and semantically modified scales, ranging from 82.2 in Form 1 (between 4C-FGD and 4C-FSGD) to 193.68 in Form 4 (between 5C-FGD and 5C-FSGD). Among the eight scales/rubrics, 3C-FSD showed the smallest AIC and BIC values, followed by 3C, 3-category scale, original rubric. While AIC and BIC values favor 3C-FSD and 3C, four other models (i.e., 4C-FSGD, 4C, 4C-FSD, and 5C-FGD) also demonstrated good fit from the values of the other indices, exclusion of the four models is not recommended. However, the frequency-based semantic judgement has a clear advantage over the quality-based counterpart and should be considered a viable alternative.

Table 7.7 Model Fit of the Eight Ordinal Scales/Rubrics (Best and Worst Values Across the Four Forms), (N=360 = 90 × 4 forms)

Scale/	Paraphrase as a combined category				Paraphrase as an independent category			
Rubric	3C	3C-FSD	4C-FGD	4C-FSGD	4C	4C-FSD	5C-FGD	5C-FSGD
Form 1								
AIC	1876.92	1629.14	2342.52	2260.32	2279.58	2036.95	2744.65	2660.93
BIC	1966.92	1719.13	2462.51	2380.31	2397.07	2154.44	2892.14	2808.42
Form 2								
AIC	1869.06	1599.59	2356.24	2214.59	2306.09	2024.93	2782.08	2638.33
BIC	1959.05	1689.58	2476.23	2334.58	2426.09	2144.92	2932.07	2788.32
Form 3								
AIC	1836.42	1607.61	2359.02	2243.16	2157.93	1943.48	2691.02	2576.42
BIC	1926.42	1695.10	2479.01	2360.65	2275.42	2058.47	2838.51	2721.41
Form 4								
AIC	2514.63	1624.83	2514.63	2327.87	2316.39	1988.15	2880.95	2687.27
BIC	2634.63	1714.83	2634.63	2447.86	2436.38	2108.14	3030.94	2837.26

For the test-level evaluation, marginal reliability was also examined. Table 7.8 shows that reliability was increased when either frequency-based semantic or grammatical judgement was used across all forms. Interestingly, the differences between the rubrics that modified the criteria of either grammar or semantic judgement only were marginal, but the models run with the rubrics that modified both criteria performed best. In summary, the model fit analysis, combined with test reliability, resulted in preference for the 3-cateogry FSD or FSGD rubrics.

Table 7.8 The Range of Marginal Reliability of the Eight Ordinal Scales/Rubrics

Scale	Paraphrase as a combined category				Paraphrase as an independent category			
/Rubric	3C	3C-FSD	4C-FGD	4C-FSGD	4C	4C-FSD	5C-FGD	5C-FSGD
Form 1	0.86	0.88	0.88	0.91*	0.87	0.89	0.89	0.91*
Form 2	0.89	0.92	0.90	0.93*	0.90	0.92	0.91	0.93*
Form 3	0.90	0.91	0.91	0.92	0.91	0.92	0.92	0.93*
Form 4	0.86	0.89	0.89	0.91	0.87	0.89	0.89	0.92*
Lowest	0.86	0.88	0.88	0.91	0.87	0.89	0.89	0.92

Note: * the highest reliability of each form

7.2.2 Item Parameters

Item discrimination and item difficulty of the eight models were compared to examine how frequency-based semantic and grammatical judgement impacted the item parameters. Table 7.9 presents the ranges of item discrimination and difficulty of the eight scales/rubrics, as well as the frequency of parameters with standard errors larger than 0.5.

Overall, modification of criteria made some changes in item discrimination between some scales. From the correlations described in Appendix 7.1, the most dissimilar pair was 3C and 5C-FSGD (r = 0.6), which are different in both rating criteria (i.e., semantic and grammatical judgements) and the use of the paraphrase category. The pairs with the original scale and the one with the refined minor error categories showed the highest and almost perfect correlations, 0.96 between 3C and 4C-FGD and between 4C and 5C-FGD.

Figure 7.1 provides a graphic comparison of the item discrimination values among the four scales. The lines reflect the smoothed values. In general, modifying the criteria of semantic and grammatical judgments from quality-based to frequency-based approach improved item discrimination. In Figure 7.1, the most modified scale/rubric, 4C-FSGD, generated most discriminating set of items while the least modified scale, 4C, resulted in the lowest set. The smoothed lines of the item discrimination values show that item discrimination was highest when scales use both the collapsed paraphrase category and frequency-based semantic rating criteria (i.e., 4C-FSGD, 3C-FSD), followed by the scales that made either modification (i.e., 3C, 4C-FSD).

The improvement was more obvious for the lower values. Table 7.9 also reveal that the maximum values were improved only with 3C-FSD and 4C-FSD, which modified semantic judgement only. Because the maximum item discrimination of each scale/rubric is 2.96 or higher, all of which are larger than 1.70, the threshold of high discrimination (Baker & Kim, 2017), the chances in the minimum values were more important. Increases were found in all modified scales/rubrics, but the frequency-based semantic approach made slightly larger improvement than modifying grammatical criteria. The lowest discriminating item on the original rubrics, 3C and 4C, was Item 37. The discrimination of Item 37 (i.e., *As you can see on the course schedule, we will not have a final exam for this course*.) was increased from 0.84 (in 3C) to 1.65 (in 3C-FSD) and 2.07 (in 4C-FSGD), and from 0.92 (in 4C) to 1.79 (4C-FSD) and 2.20 (in FSGD). With these changes, the least discriminating item, Item 7 (i.e., *Earning money is the main reason for students to get a job*), approached high discrimination in 3C-FSD (a = 1.26) and in 4C-FSGD (a = 1.24).

Similar but smaller changes were found among the scales that have both exact repetition and paraphrase categories.

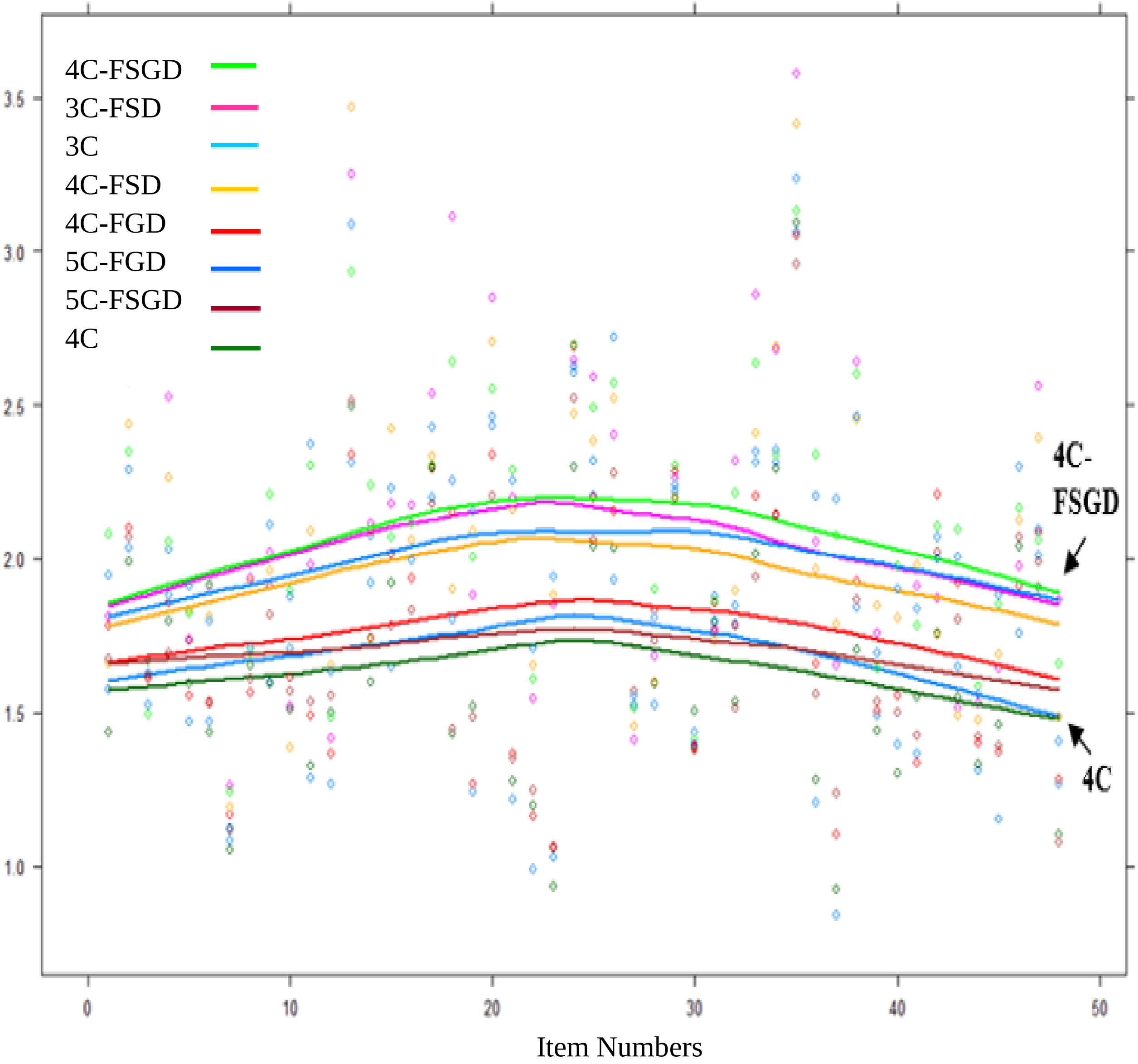


Figure 7.1 Item Discrimination by Eight Scales/Rubrics

Table 7.9 The Range of Item Parameters of the Eight Ordinal Scales/Rubrics

Scale/	Paraphrase as a combined category				Paraphrase as an independent category			
Rubric	3C	3C-FSD	4C-FGD	4C-FSGD	4C	4C-FSD	5C-FGD	5C-FSGD
Range of item discrimination, *a* (Item)								
Min.	0.84 (#37)	1.26 (#7)	1.06 (#23)	1.24 (#7)	0.92 (#37)	1.19 (#7)	1.06 (#23)	1.12 (#7)
Max.	3.23 (#35)	3.58 (#35)	3.05 (#35)	3.13 (#35)	3.09 (#35)	3.47 (#13)	2.96 (#35)	3.08 (#13)
Range of item difficulty, *b* (Item)								
$b_{overall}$								
Min.	-1.21 (#18)	-1.53 (#7)	-1.38 (#18)	-1.67 (#7)	-0.97 (#18)	-1.00 (#18)	-1.18 (#18)	-1.30 (#7)
Max.	1.71 (#11)	1.38 (#11)	1.71 (#10)	1.37 (#11)	2.43 (#11)	2.94 (#11)	2.10 (#11)	1.70 (#11)
$b_{paraphrase}$								
Min.	-0.60 (#18)	-0.51 (#18)	-0.60 (#18)	-0.51 (#18)	-0.66 (#18)	-0.51 (#18)	-0.62 (#18)	-0.50 (#18)
Max.	3.18 (#11)	2.66 (#11)	2.90 (#11)	2.51 (#11)	3.08 (#11)	2.75 (#10)	2.81 (#11)	2.43 (#11)
Number of parameters with standard error > 0.5								
a	7	13	1	3	2	8	2	2
b	5	5	3	2	18 (12*)	12 (8*)	14 (11*)	8 (5*)

Note: * Number of item difficulty parameters of exact repetition with SE > 0.5

The modifications generally made the items easier in terms of overall difficulty to some degree. In Table 7.9, the minimum was lowered, particularly on the paraphrase-collapsed rubrics with frequency-based semantic judgement. On the scales with the independent paraphrase category, however, the frequency-based approach to both semantic and grammatical judgement (i.e., 5C-FSGD) was found to be easiest. To compare the changes by different scales/rubrics across the eight options, the thresholds of the paraphrase category were compared, and a similar pattern was found. Appendix 7.2 showed that correlations of the category item difficulty among the eight scales were very high, 0.98 or higher. Despite the overall strong relationships, modified scales/rubrics somewhat lowered the thresholds of the paraphrase category. Frequency-based semantic judgement (in 3C-FSD and 4C-FSD) impacted more than grammatical judgement (in 4C-FGD and 5C-FGD), but the combined approach (in 4C-FSGD and 5C-FSGD) led to the lowest threshold. The lowered item difficulty is likely to increase item discrimination because too difficult items are discriminating than adequately difficult items.

Interestingly, however, examining SE of the item discrimination and difficulty parameters provided somewhat different information. Using frequency-based semantic judgement led to more items with SE beyond the acceptable range (SE > 0.5), from 7 (in 3C) to 13 (in 3C-FSD), and from 2 (in 4C) to 8 (in 4C-FSD), but modification on grammatical judgement alone (i.e., 4C-FGD, 5C-FGD) or combined uses (i.e., 4C-FSGD, 5C-FSGD) reduced the number of discrimination parameters lacking preciseness. Similar patterns were found in the preciseness of the item difficulty parameters. One noticeable aspect is that the number of item difficulty parameters of a preciseness concern was much larger on the rubrics with the independent categories of paraphrase and exact repetition (i.e., 4C, 4C-FSD, 4C-FGD, 4C-FSGD). When closely reviewed, the majority of the item parameters beyond the acceptable precision were the item difficulty parameters of the exact repetition category, which aligned with the results from the larger samples in the previous chapters. When both appropriate preciseness and item parameter values considered, 4C-FSGD, which is the scale/rubric that collapsed paraphrase into exact repetition and used frequency-based semantic and grammatical judgement, was found to be most optimal. The 4C-FSGD model also resulted in the second-best overall model fit, following the 3C-FSD model.

7.2.3 Misfit

Model fit analysis included both item and person levels. Table 7.10 and 7.11 shows the number of misfitting items and examinees on each scale/rubric across the four forms.

According to Table 7.10, the FSGD model was found to perform best, having no item misfit at the conventional significance level ($p < 0.05$), which used a frequency-based approach to both semantic and grammatical judgement and collapsed paraphrase into exact repetition. The 3C-FS also performed well and resulted in only one item misfit, which combined the two highest categories, but did not make semantic or grammatical modification to the original rating rubric. Thus, frequency-based modification of rating criteria alone did not decrease the number of item misfits. Rather, combining exact repetition and paraphrase alone positively influenced the item misfit overall, particularly from the comparison between 3C and 4C, and between 4C-FSGD and 5C-FSGD. Note that the 4C-FSD model has one item (i.e., Item 18) flagged even after a Bonferroni adjustment. The effect of combining the two highest categories was best when used with both frequency-based modifications, as shown in the comparison between 3C and 4C-FSGD.

Table 7.10 The Number of Misfitting Items on the Eight Scales/Rubrics Across Four Forms

Scale/	Number of misfitting items (S-X^2, p < 0.05)						Bonferroni
Rubric	Form 1	Form 2	Form 3	Form 4	Total	Items	Corrected
Paraphrase as a combined category							
3C	0	0	0	1	1	Q46	
3C-FSD	1	1	0	1	3	Q8, Q21, Q45	
4C-FGD	1	1	1	1	4	Q11, Q18, Q33, Q43	
4C-FSGD	0	0	0	0	0	NA	
Paraphrase as an independent category							
4C	1	0	1	1	3	Q5, Q31, Q38	
4C-FSD	0	1	0	2	3	Q18*, Q43, Q45	*p < 0.002
5C-FGD	1	2	1	2	6	Q6, Q16, Q18, Q39, Q43	
5C-FSGD	2	0	2	1	5	Q3, Q6, Q31, Q36, Q45	

Note: Bonferroni corrected : p < 0.0042

Table 7.11 The Number of Misfitting Examinees on the Eight Scales/Rubrics Across Four Forms

Scale/ Rubric	Number of misfitting examinees Zh > | ± 2.0|					Zh > | ± 3.0|	
	Form 1	Form 2	Form 3	Form 4	Total	Total	Examinee ID (Zh)
Paraphrase as a combined category							
3C	4	3	2	3	12	0	-
3C-FSD	0	1	2	4	7	0	-
4C-FGD	0	1	1	0	2	0	-
4C-FSGD	0	0	0	0	0	0	-
Paraphrase as an independent category							
4C	5	3	2	4	14	2	Form 1 (F1-65, Zh = -3.22), Form 2 (F2-62, Zh = -3.08)
4C-FSD	2	1	0	3	6	0	-
5C-FGD	1	3	0	3	7	0	-
5C-FSGD	2	1	0	3	6	0	-

Note: *****Zh > | ± 3.0|

Similar to the findings from the item misfit analysis, the 4C-FSGD model performed best in person misfit, as well. Table 7.11 shows that no examinee was flagged in that all showed smaller absolute Zh values than 2.0. Also, collapsing paraphrase and exact repetition reduced the number of misfitting examinees overall, particularly the pairs that include frequency-based grammatical judgement, 4C-FGD versus 5C-FGD, and 4C-FSGD versus 5C-FSGD. One difference is that the frequency-based modifications reduced the number of misfitting examinees across the eight scales/rubrics. Notably, the 4C model, which has an independent paraphrase category, two severe cases of person misfit (F1-65, Zh = -3.22; F2-62, Zh = -3.08) but the degree of misfit became less severe when either or both of the frequency-based semantic and grammatical rating criteria were used.

In summary, frequency-based modifications and collapsed category of paraphrase improved the performance in one or more fit indices, but the best-performing ordinal scale/rubric was 4C-FSGD, which used both frequency-based semantic and grammatical judgement with a combined category of paraphrase and exact repetition.

7.3 Comparison of the EI Accuracy Scales: Binary versus Ordinal Scale (RQ 4.3)

This section reports the results of the comparison between ordinal and binary scale. The ordinal scale compared with the binary scale was the 4C-FSGD model, which the previous section found best-performing. Table 7.12 shows the differences in the scoring methods between the two scales.

Table 7.12 The Scales / Rubrics Compared in Section 7.3

Scale/Rubric	Number of category	Use of the paraphrase category	Rating Criteria of Deviation / Error	
			semantic	grammatical
4C-FSGD	4	collapsed	frequency-based	frequency-based
binary scale	2	collapsed	yes-no	yes-no

7.3.1 Model Fit

In order to evaluate overall performance, model fit and reliability of the two scales were examined. Table 7.13 shows the model fit range of the best-performing ordinal scale and the binary scale. When comparing the goodness of fit, although the ordinal model fit slightly better in SRMR, both models indicated good fit. Across the four forms, RMSEA values were 0.05 or lower, TLI and CFI 0.95 or higher, and p-values were not significant. AIC and BIC definitely favored the binary scale, because the number of parameters were only the half of the ordinal scale. The result is not congruent with literature, which found ordinal scales performed better than binary scales, because the model fit of the binary scale is as good as the best-performing ordinal scale and better than some ordinal scales compared in Section 7.2.

Table 7.13 The Range of the Model Fit of the Best-Performing Ordinal and Binary Scales across the Four Forms, (N=360 = 90 × 4 forms)

Scales/Rubric	Ordinal (4C-FSGD)		Binary	
	Best fit	Least fit	Best fit	Least fit
Form	1	4	2	1
M^2	13.94	37.39	48.30	63.88
Parameters (*df*)	48 (30)	48 (30)	24 (54)	24 (54)
p (M^2)	0.99	0.17	0.69	0.17
RMSEA (M^2)	0	0.05	0	0.05
SRMR (M^2)	0.05	0.06	0.07	0.07
TLI	1.22	0.96	1.01	0.97
CFI	1.00	0.97	1.00	0.98
LLV	-1082.16	-1115.93	-470.87	-502.40
AIC	2260.32	2327.87	989.73	1052.80
BIC	2380.31	2447.86	1049.73	1112.80

Notes: LLV: log-likelihood value

In the case of preciseness, however, the binary scale was rather problematic. According to Table 7.14, the marginal reliability of the binary scale ranged from 0.75 to 0.81, which is the borderline of the acceptable reliability for a high-stakes exam or below. The reliability is fairly low compared with the best-performing ordinal scale's lowest precision (r = 0.91), and lower than the lowest precision among the eight ordinal models (r = 0.86). Thus, when both model fit and reliability considered, the ordinal scale performed better.

Table 7.14 The Range of Marginal Reliability of the Best-Performing Ordinal Scale and the Binary Scale

Form	Ordinal (4C-FSGD)	Binary
Form 1	0.91	0.80
Form 2	0.93	0.81
Form 3	0.92	0.80
Form 4	0.91	0.75
Lowest	0.91	0.75

Note: * the highest reliability of each form

7.3.2 Item Parameters

For the item level comparison, item parameter and SE values of the two scales were examined. Table 7.15 shows the minimum and maximum values of the parameters and the number of parameters with SE beyond the generally acceptable precision.

Overall, the range was wider on the binary scale for both discrimination and difficulty parameters. Importantly, the minimum discrimination value of the binary scale was fairly lower than the minimum on the ordinal scale. The least discriminating item on the ordinary scale, Item 7 (a = 1.24), actually improved its discrimination on the binary scale (a = 1.33), but the decreases in other four items (i.e., Item 9, Item 37, Item 39, and Item 45) were large, which caused the minimum lower and the range wider. Particularly, the change in the least discriminating item, Item 45 (i.e., *When you take courses here, attendance often counts as a part of the final grades*), was noticeable. The discrimination of Item 45 was 1.85 on the ordinal scale, classified as very high discrimination (a > 1.70), but was reduced to 0.88, which was still considered moderate discrimination but approached the borderline of the low discrimination (a = 0.65), according to Baker and Kim (2017). Similarly, the most highly discriminating item on the ordinal scale, Item 35 (a = 3.13) went through only a slight increase on the binary scale (a = 3.64), both of which are considered very high discrimination. However, five other items (i.e., Item 4, Item 10, Item 17, Item 20, Item 25, Item 33) underwent considerable increase in their discrimination values. Item 10 (i.e., *You can tell me what questions you have on the final project during my office hours*), the most discriminating item, displayed a sharp increase in its discrimination value by 3.04 from 1.89 on the ordinal scale to 4.93 on the binary scale, although both values are considered very high discrimination.

The range of item difficulty or the threshold of the paraphrase category was also wider on the binary category, but the difference was small and limited to few items. While the lowest item difficulty values of the paraphrase category were similar on both scales, the maximum value was considerably increased. However, unlike the distribution of item discrimination, the change was due to one item, Item 45. The item difficulty of Item 45 was also shifted most from 0.51 on the ordinal scale to 3.18 on the binary scale. Most item difficulty were similar, and there were eight items that underwent a change of | ± 0.2 |, decreases in two items (i.e., Item 10, Item 11) and increases in six items (i.e., Item 9, Item 13, Item 23, Item 37, Item 39, and Item 45). Among which the changes in two items, Item 39 (Δ = 0.58) and Item 45 (Δ = 1.22), were considerable.

Table 7.15 The Range of the Item Parameters of the Best-Performing Ordinal Scale and Binary Scale

Scale/Rubric	Ordinal (4C-FSGD)	Binary
Item discrimination, *a* (Item)		
Min.	1.24 (#7)	0.88 (#45)
Max.	3.13 (#35)	4.93 (#10)
Item difficulty, $b_{paraphrase}$ (Item)		
Min.	-0.51 (#18)	-0.59 (#18)
Max.	2.51 (#11)	3.18 (#45)
Standard error > 0.5 (freq.)		
a	3	33
b	2	5

The examination of the SE values revealed more noticeable differences, which clearly indicated better performance of the ordinal scale. The number of the discrimination parameters with a larger SE value than acceptable precision (SE > 0.5) was 33 out of 48 items, among which SEs of eight items were larger than 1.0. Item 10, the most discriminating item on the binary scale (a = 4.93) showed an SE of 4.37, which drastically degrades the confidence in the item parameter. On the other hand, only three items were flagged on the ordinal scale. Estimation of the item difficulty parameters on the binary scale was much more precise than of item discrimination. Five items were found to have SE larger than 0.5, which was far smaller than the number of flagged item discrimination, although the number is still slightly larger than the number of flagged items difficulty parameters on the ordinal scale. The most difficult item on the binary scale, Item 45, was most imprecise in that its SE is 1.51. The precision issue serves as strong evidence for superior performance of the ordinal scale despite the binary scale's higher item discrimination.

7.3.3 Misfit

Misfit analysis also supports the outperformance of the ordinal scale. Table 7.16 and Table 7.17 present the number of misfitting items and examinees in each form on ordinal and binary scales, respectively. According to Table 7.16, the binary scale detected two misfitting items at the conventional significance level ($p < 0.05$), Item 8 (i.e., *If you record your lectures, you can revise your class notes later.*) and Item 16 (i.e., *The wonderful thing about English teachers is that they know their students quite well*). Thus, ordinal scale fit better to some degree. After the Bonferroni correction scale, neither scale had a misfitting item.

Table 7.16 The Number of Misfitting Items on the Ordinal and Binary Scales

Scale/Rubric	Number of misfitting items (S-X^2, $p < 0.05$)						Bonferroni Corrected
	Form 1	Form 2	Form 3	Form 4	Total	Items	($p < 0.0042$)
Ordinal (4C-FSGD)	0	0	0	0	0	-	-
Binary	1	1	0	0	2	Q8, Q16	-

Table 7.17 The Number of Misfitting Examinees on the Ordinal and Binary Scales

Scale/ Rubric	Number of misfitting examinees						
	Zh > | ± 2.0|					Zh > | ± 3.0|	
	Form 1	Form 2	Form 3	Form 4	Total	Total	Examinee ID (Zh)
Ordinal (4C-FSGD)	0	0	0	0	0	0	-
Binary	1	2	2	2	7	0	-

The binary scale also identified seven examinees with flagged performance patterns. While the number of misfitting items and examinees on the binary scale were not the worst compared with other ordinal scales, was fairly worse than the best-performing ordinal scale because no item or examinee was flagged on the ordinal scale. Thus, the results of the misfit comparison also support the ordinal scale, in line with the analyses of the model fit, item parameters, and precision.

7.4 Discussion and Conclusion

Chapter 7 was concerned about scale revision and validation. First, I described how the alternative scales and rubrics were developed and addressed the issues identified from the qualitative analysis in Chapter 6 (RQ4.1). Next, I presented the results of the comparison among the eight ordinal scales (RQ4.2) and between the best-performing ordinal scale and the binary scale (RQ4.3).

For RQ 4.1, three alternative approaches were proposed by modifying the grammatical and semantic rating criteria of the original rubric: frequency-based rating to either grammatical or semantic performance, or both. The three approaches were controlled for the combined or independent use of paraphrase and exact repetition, which resulted in six alternative ordinal scales/rubrics. In addition, a binary scale was proposed for pragmatic benefits, which rates whether or not exactly repeat/appropriately paraphrase.

As a response to RQ4.2, among the six alternative ordinal scales/rubrics and two original rubrics with or without paraphrase collapsed, the most modified scale/rubric (4C-FSGD) was found to perform best, which is the four-category scale that employs frequency-based semantic and grammatical rating and a combined category of paraphrase and exact repetition. The outperformance of the modified scale/rubric brings attention to three points worth noting. First, in line with the analysis of the larger sample in the previous chapters, the combined use of paraphrase and exact repetition performs better. This alignment adds evidence for the potential construct validity issue of distinguishing exact repetition and paraphrase. Evidence related to other criterions, for example, comparison with other speaking task or test performance of examinees who paraphrased and exactly repeated, would be useful to further clarify the construct validity issue. Also, process-based research, such as a think aloud protocol or retrospective cognitive interviews, would provide clearer picture on why examinees happen to or choose to paraphrase a prompt.

Second, the combined use of frequency-based grammatical and semantic rating performed better than either was used alone or none. The improved performance likely suggests that the modified rubric reflects the language construct that EI measures better. The quantitative approach to semantic rating considers both lexical sophistication and grammatical complexity of the responses while the original approach actually focuses on core meaning only. One-level more fine-grained classification of grammatical performance include not only the degree of errors but also the quantity of errors while the original approach only considers the former. The better performance of the modified rating approaches indicate that linguistic complexity/sophistication are important for language proficiency. It would be a useful addition to examine rating process of how raters apply the modified and original scoring method in relation to assessing the construct measured. Particularly considering that the combined rubric might affect cognitive load on the raters, compared to the existing rubric, it would be necessary to obtain evidence that confirms feasibility of this new approach if the new method is to be used with typical rater pool, rather than automatic scoring.

Third, importantly, note that using both frequency-based ratings, rather than only modification alone, was most effective. Using frequency-based semantic rating alone produced great model fit but precision was decreased to some degree while frequency-based grammatical rating criteria worked the other way to some degree. The outperformance of using both approaches together (with a combined category of paraphrase and exact repetition) supports L2 English language proficiency is a multi-faceted construct that includes both complexity and accuracy. In other words, although it is possible and useful to analytically apply fine-grained grammatical and semantic accuracy, when one aspect is over-emphasized, scoring might not reflect the construct appropriately. L2 users have different profiles, which means different strengths and weaknesses, and taking both modifications into account allows assess the wide range of profiles more accurately. Examination of relationships between fine-grained linguistic features of prompts and scoring would further elucidate the role of linguistic complexity in assessing language proficiency in EI. In a similar sense, future studies are strongly encouraged to include fluency for more complete view of language proficiency measured by EI, because fluency not only is correlated with accuracy and complexity but also assesses a unique aspect of language proficiency.

To answer the final question, RQ 4.3, I further compared the best-performing ordinal scale with the binary scale and found superior performance of the ordinal scale. The finding partially aligns with previous study that found that ordinal scale better distinguish L2 proficiency than binary scales (Yan et al., 2016). Although the binary scale underperformed the best-performing ordinal scale and revealed an obvious precision issue, the model fit was comparable to some degree, and overall performance was better than or as good as some ordinal scales. This difference is probably due to the different rating criteria of the binary scale in the current study. In previous studies, binary scales generally rated whether or not to exactly repeat the given prompt while the current study's binary scale distinguished appropriate paraphrase and above from below, that is, a combined category of exact repetition and paraphrase. This indicated that rating criteria itself is more or as important as the scale type as well as emphasizes the potential appropriateness of combining exact repetition and paraphrase. On the other hand, the precision issue that binary scale encountered can imply that the inconsistency issues between appropriate and non-appropriate paraphrase might degrade precision. The precision might have been higher if simplified paraphrase and omission of semantically non-essential parts had been consistently penalized. Future research is invited to clarify this possibility.

PHASE II. BEYOND MEASURING ACCURACY: RELATIONSHIP BETWEEN EI PROMT COMPLEXITY AND EI MEASURMENT OF ACCURACY

Phase II presents findings on the relationships between accuracy measurement qualities and prompt complexity. Using MLM and RF analyses, Chapter 8 and Chapter 9 report on the results on how prompt length and linguistic features of EI items are related to two measurement qualities of EI items (i.e., item difficulty, item discrimination) respectively. In Phase I, item difficulty was found to be a crucial contributor to the category adequacy of the current test. Thus, examining the impact of such main prompt characteristics as sentence length and linguistic features to item difficulty make important implications for improving the adequacy of category usage as well as the overall measurement quality and test effectiveness. In addition, Phase II elaborates the results on the extent to which prompt length and linguistic features influence EI item discrimination, another important parameter of item measurement. The findings offer useful insight into item and test quality control in general, while providing specific information so test/item developers apply prompts of different lengths and linguistic features to adjust item difficulty and discrimination together.

CHAPTER 8. IMPACT OF PROMPT LENGTH ON EI ITEM DIFFICULTY AND DISCRIMINATION

In regard to the impact of prompt length on EI item measurement qualities, Chapter 8 answered the following research questions:

RQ 5.1 To what extent, does prompt length (i.e., number of syllables, number of words) impact item difficulty across eleven different scales/ scoring methods? (Section 8.1)

RQ 5.2 (a) To what extent, does prompt length (i.e., number of syllables) impact item discrimination across ten different ordinal scales/scoring methods? (b) Which characteristics of scale modification are associated with the relationship? (Section 8.2)

Results from univariate regression and MLM analyses indicated that prompt length significantly impacted item difficulty and item discrimination across the scales/scoring methods. The magnitude of the impact was much smaller on item discrimination than on item difficulty while the variance explained was small for both parameters. Details are provided as follows.

8.1 Impact of Prompt Length on Item Difficulty (RQ 5.1)

To examine the relationship between prompt length and item difficulty (at the paraphrase level), univariate regression was selected over multi-level modeling, although the data had a nested structure, where 48 items or item difficulty values (at the paraphrase level) were nested within each of the eleven scales and rubrics, which were one binary and eight ordinal scales (i.e., 3C, 3C-FSD, 4C-FGD, 4C-FSGD, 4C, 4C-FSD, 5C-FGD, and 5C-FSGD) with the small sample (N = 360 in total) and two ordinal scales (i.e., 3C and 4C) with the entire sample (N = 799). The choice was made because the null model of two-level multilevel modeling with scale/scoring methods as a group variable (N=11) and item difficulty values (N = 48) as an outcome variable revealed that the intraclass correlation (ICC) was almost zero. The ICC value was obtained from the following equation based on the null model with a random intercept and no predictors:

$$ICC = \frac{group\ variance\ (0.0000000004)}{group\ variance\ (0.0000000004) + residuals\ (0.5782538)}$$

The ratio of variance from the nesting groups (i.e., eleven scales/scoring methods) to the total variance was extremely small, which means individual values within each group (i.e., item difficulty values) are independent from the scale/scoring methods. The independence assures the use of a simpler statistical approach, a univariate regression to examine the item difficulty across the groups.

Table 8.1 presents the results of six regression analyses, A1 to B3, where prompt length (i.e., the number of syllables or word count) predicted item difficulty (i.e., the accumulative threshold of the paraphrase category (or higher)). Regression model A1 and B1 analyzed the item difficulty obtained from all scales/methods and sample options (N = 48 × 11 = 528). Model A2 and B2 employed item difficulty of the scores rated by the original criteria on the 3-category and 4-category EI scales with the entire sample (N = 779), which were used from Chapter 4 to 6, while Model A3 and C3 is based on item difficulty calibrated by using the scores from the 4C-FSGD scale/rubric with the smaller sample (N = 360), which Chapter 7 found best-performing. These subsets were additionally analyzed because the item difficulty from the original scoring rubric except for the highest/lowest categories collapsed used for a entire sample would give useful information to understand the current scoring. Also, the information abased on item difficulty from the best-performing model would be instrumental for the future development and revision of scales and rating rubrics.

Table 8.1 Results of Four Single Linear Regression Analyses for Prompt Length as a Predictor of Item Difficulty* and Item Discrimination Across Scales

Model	Scale	Sample	Predictor Variable	R^2	*Adj*. R^2	*B*	*t* (*df*)	*P*
A1	All**	all	number of syllables	0.27	0.27	0.17	13.86 (526)	<0.0001
A2	3C & 4C	large	number of syllables	0.25	0.24	0.18	5.63 (94)	<0.0001
A3	4C- FSGD	small	number of syllables	0.28	0.26	0.18	4.19 (46)	0.0001
B1	All**	all	word count	0.09	0.09	0.10	7.336 (526)	<0.0001
B2	3C & 4C	large	word count	0.11	0.10	0.11	3.36 (94)	0.001
B3	4C- FSGD	small	word count	0.09	0.07	0.10	2.16 (46)	0.036

Note. item difficulty* = the accumulative threshold of the paraphrase category or higher (b_2); All** = item difficulty values from eleven scales

According to Table 8.1, sentence length positively and significantly impacted item difficulty, in line with previous research (Bley-Vroman & Chaudron, 1994; Campfield, 2017; Graham et al, 2010; Yan et al., 2016). Interestingly, unlike some previous studies (e.g., Campfield, 2017), the current analysis found that the number of syllables predicted item difficulty better than the number of words did (from the comparison of Model A1 versus B1, A2 versus B2, and A3 versus B3). The positive relationships between the number of syllables/word count and item difficulty and outperformance of the number of syllables were consistent across the scales/scoring methods. The minimal difference was expected from the low ICC and from the high correlations of item difficulty from the 4C-FSGD scale/rubric with those of the original rubric on both 3- and 4-category scales with both small and entire samples (r = 0.95 or higher).

Importantly, however, the amount of variation in item difficulty explained by prompt length was relatively small, maximum 27% by the number of syllables (in Model A1) and 10% by word count (in Model B2). One reason is that the range of the sentence length is restricted in the current study, from 14 syllables (and 9 words) to 22 syllables (and 18 words). If very short or longer items are included, the results would be different. However, more important consideration should be that sentence length is not the only factor that affects item difficulty of EI but rather one of the important predictors. The relatively small variation explained, although not small as a single predictor, makes sense because the construct that EI measures, L2 proficiency, is not exclusively about sentence length, which is related to one aspect of syntactic complexity.

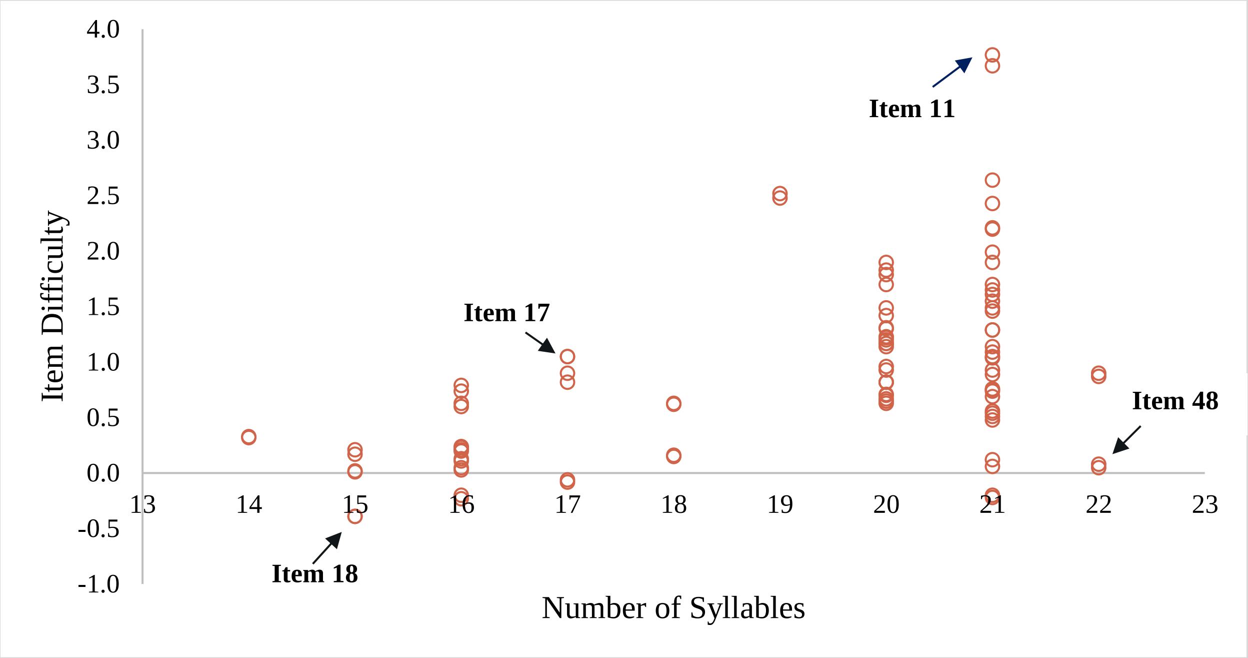


Figure 8.1 Item Difficulty by the EI Prompt Length (Number of Syllables) in Model A1

Figure 8.1 presents the distribution of item difficulty by the number of syllables in a prompt and shows variation in item difficulty within a prompt of the same syllable length (in Model A2). The variation was generally larger in longer sentences. While the number of syllables of highly difficult items was large, items with low item difficulty varied in the number of syllables. For example, items with the threshold value of the paraphrase category 1.5 or higher has 19 or more syllables, but item difficulty of zero was found with items across the range of sentence length. Thus, longer prompts were not always more difficult than shorter prompts. Among the 48 items, the most difficult item was Item 11, length of which was the second longest (i.e., 21 syllables), while the easiest item was Item 18, whose length was the second shortest. The length-difficulty relationships in these items are well aligned with the literature—the longer the more difficult—there were some items that were not, such as Item 17 and Item 48. Item 48 was the longest item but found to be almost the easiest, and Item 17 was somewhat shorter, but its item difficulty was not low.

The discrepancies lie in lexico-grammatical complexity. From the prompts provided below, Item 11 requires a highly sophisticated syntactic structure by using a gerund subject, present perfect tense combined with passive voice and third person singular (*has been shown*) and complex object (of *help*), as well as a phrasal verb (*succeed in*). On the other hand, the long sentence length of Item 48 primarily relies on connecting simple clauses with most commonly used lexis, such as *You should talk to*, *if you are not sure*, and *courses to take*. Similarly, Item 17 uses a somewhat complex object that consists of noun and prepositional phrases (*a topic for your final project by midterm*) as well as a three-word collocation (comp up with), whereas Item 18 uses most basic lexis (*looks like, have morning classes*) and an omitted *that* complement connecting a simple clause, which is most commonly used by first-year L2 users in college (Shin, 2021). Thus, while adjusting item difficulty to increase the adequacy of the independent use of paraphrase and exact repetition, it is important to consider both sentence length and lexical and syntactic complexity. Also, the difficulty examined in this section is at the paraphrase level. Analysis based on item difficulty of other levels is expected to lead to (slightly) different results. Specific information on which linguistic features are related to item difficulty across different item category levels and scales is provided in Section 8.3.

Item 11 *Taking a part-time job on campus has been shown to help students succeed in college* (b_2 = 3.67, 21 syllables)

Item 17 *You should come up with a topic for your final project by midterm.* (b_2 = 1.05, 17 syllables)

Item 18 *It looks like I only have morning classes this semester.* (b_2 = 0.35, 15 syllables)

Item 48 *You should talk to your advisor if you are not sure what courses to take next semester.* (b_2 = 0.08, 22 syllables)

8.2 Impact of Prompt Length on Item Discrimination (RQ5.2)

In Chapter 7, modification of rating criteria made some differences in item discrimination across the scales/ scoring methods. Thus, multi-modeling analysis was considered to examine the relationship between prompt length and item discrimination, which has ten different ordinal scales/scoring methods as a group variable. The binary scale was excluded from the examination of the associations because the estimated item discrimination parameters of the majority of the items were beyond the acceptable range of precision (SE > 0.5) in Chapter 7. For sentence length, only the number of syllables was examined because syllable-based prompt length was a more effective measure of item difficulty in Section 8.1.

Before conducting the MLM analysis, I plotted the relationships between prompt length and item discrimination of the 48 EI items across the ten scales/scoring methods in Figure 8.2. Overall, there were some positive linear relationships between prompt length and item discrimination across the scale groups, which indicated that it is reasonable to add prompt length as a predictor of item discrimination to a model.

In line with the findings in Chapter 7, however, item discrimination was inconsistent among the ten scale groups, particularly within each unit of sentence length, including the intercepts. The differences of item discrimination (outcome variable) between scales invite to use MLM so the variance due to the nested structure can be considered in the statistical model. Also, note that the variation after controlling for prompt length within each scale indicates exploring other predictors, such as prompt linguistic features, which was discussed in Section 8.4.

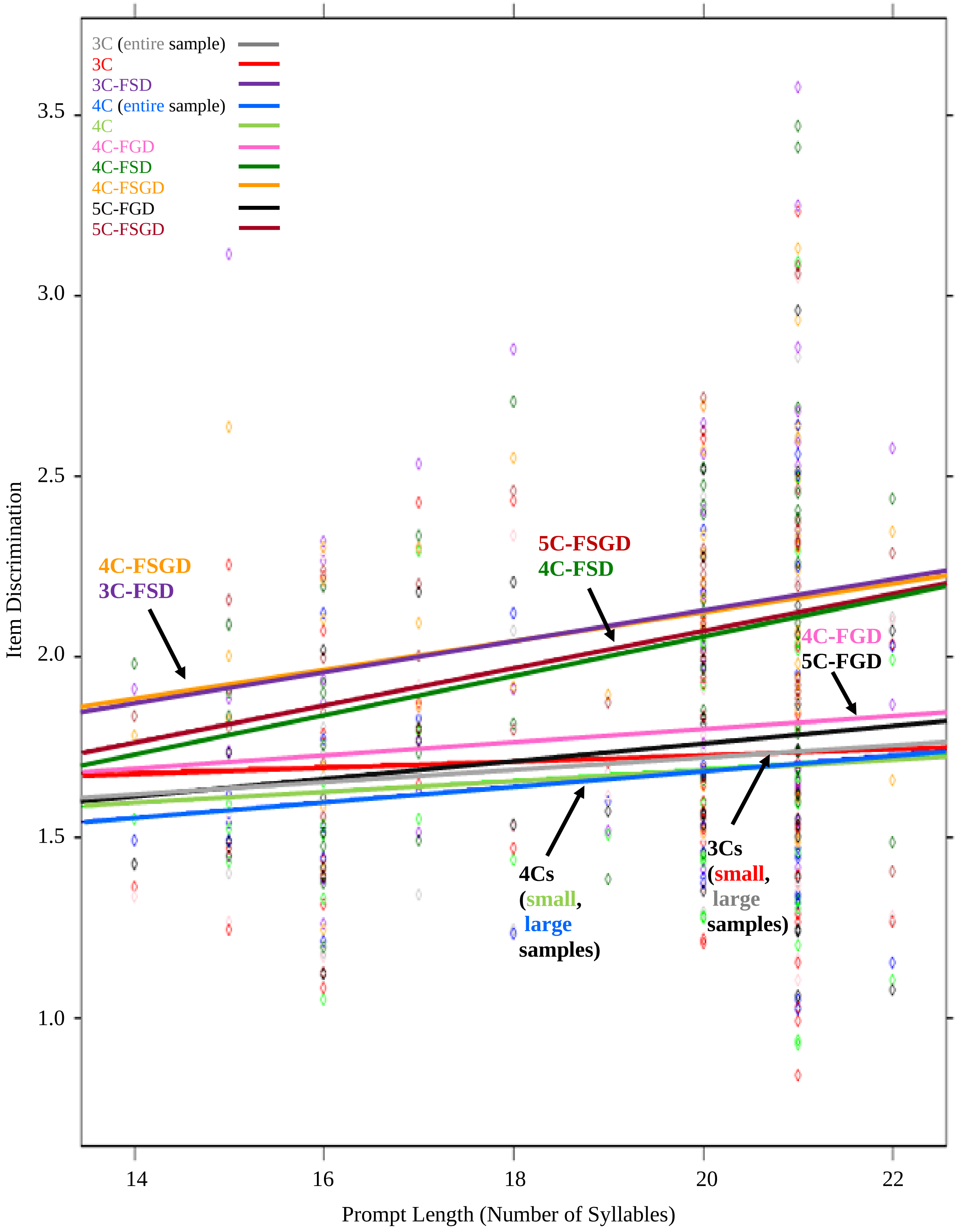


Figure 8.2 Item Discrimination by Prompt Length Across Ten Ordinal Scales

The closer examination of the plot proposed exploring three aspects of modified rating scales as predictors of the group variation in item discrimination: (1) frequency-based semantic rating (FS), (2) frequency-based grammar rating (FG) of minor grammatical errors, and (3) collapsed category of paraphrase (CCP) into exact repetition. First, the four scale groups that employed FS (i.e., 4C-FSGD, 3C-FSD, 5C-FSGD, and 4C-FSD) behaved somewhat differently from the rest. These scales showed higher intercepts and steeper slopes, which suggests adding FS as a Level-2 predictor and inter-level interaction. Second, among the four scales using FS, the scales that used CCP (i.e., 4C-FSGD and 3C-FSD) had higher intercepts than those without CCP (i.e., 5C-FSGD and 4C-FSD). Also, among the four lowest scale groups (i.e., 3C and 4C with small and entire samples), the scales without CCP (i.e., 4Cs) had lower and the lowest intercepts, which reveals the potential role of CCP regardless of semantic or grammatical modifications. Thus, a model that enters CCP was also examined. Third, among the six scale groups without FS, the scales with FG (i.e., 4C-FGD, 5C-FGD) had higher intercepts than those without semantic or grammatical modifications (i.e., 3Cs, 4Cs), which suggests FG as a potential Level 2 predictor.

Each suggestion from the graphic observation was statistically examined via five MLM models and a Null model. First, to confirm the need to use MLM, a two-level Null model for the current data was constructed as follows:

Level-1 Model (within-scale) : $Y_{ij} = \beta_{0j} + r_{ij}$
Level-2 Model (between-scale) : $\beta_{0j} = \gamma_{00} + \mu_{0j}$

At Level 1, an item discrimination value (Y_{ij}) for a given item i by a given scale j is defined as a function of an intercept (β_{0j}) and a random component (r_{ij}), where β_{0j} is overall mean item discrimination value across all 48 items and 10 scales/scoring methods, r_{ij} is unmodeled variability between the items at Level 1. At Level 2, the intercept (β_{0j}) is a function of the regression intercept (γ_{00}) and a group-level random component (μ_{0j}), or unmodeled scale group variation. The intercept (γ_{00}), 1.5 (with an SE of 0.05), was significant ($p < 0.001$, t = 34.30, df = 470). The estimate for the between-scale variance (μ_0) was 0.025 and within-scale variance (r) was 0.187. Using these estimates, the ICC value was calculated ($0.116 \approx 0.03 / (0.03 + 0.19)$), which means 11.6 % of the total variance in item discrimination values depends on the differences between scales/scoring methods, and 88.4% is explained by within-scale (between items). The ICC value above 0.10

indicates the need to use MLM rather than regression modeling to account for the clustering (Lee, 2000) in examining the impact of prompt length on item discrimination across the ten scale groups.

With the justification, five MLM models were constructed, and the results are reported in Table 8.2. Model C1, the simplest model, examined the impact of prompt length (Level-1 predictor) on item discrimination (Level-1 outcome). The prompt length effect (γ_{10}) was 0.03, which means that item discrimination was increased by 0.03 with every increase of one syllable across the scales. The magnitude was significant (p < 0.001) but did not explain the substantial portion of the variation between items in that Model C1 reduced only 2.36 % of the within-scale variance compared to the Null model. The low variance explained calls for investigating other prompt-related sources of variance in item discrimination, such as prompt linguistic features (See section 8.4).

On the other hand, adding FS (γ_{01}) as a Level-2 predictor in addition to prompt length at Level 1 (Model C2) reduced almost all between-scale variations—100.00% rounded to the nearest hundredth—compared to the Null or Model C2. The coefficient was 0.34 and significant (p < 0.001), which means FS-based scales produced higher item discrimination values by 0.34 on average. FS as a between-scale predictor also slightly reduced the within-scale variance, 1.32% of the residuals in Model C1. However, the two models that entered FG (γ_{02} in Model C3) or CCP (γ_{02} in Model C4) in addition to prompt length and FS showed that either was significant while prompt and FS was continuously significant.

Interestingly, when an interaction between FS and prompt length was introduced (γ_{11}) in Model C5, neither the interaction nor the two preexisting predictors (i.e., prompt length, FS) was significant. The insignificant interaction was somewhat inconsistent with the graphical observation of Figure 8.2, which demonstrates the four scales with FS in general, and particularly the two scales that used FS but without CCP, had steeper slopes. From the fact that the p-value of the interaction term was smallest (p = 0.09) among the three, the coefficient of the prompt length was reduced, and the SE for FS was increased, the lack of significance in the interaction was likely due to the small number of Level-2 and small impact of sentence length on item discrimination.

Table 8.2 Results for MLM Analyses (Model C1 to C5)

	Models				
	C1	C2	C3	C4	C5
Fixed effects					
Level 1	Coefficient (standard error)				
Intercept (γ_{00})	1.29 ***	1.16 ***	1.14 ***	1.13 ***	1.38 ***
	(0.18)	(0.17)	(0.17)	(0.17)	(0.21)
Prompt length (γ_{10})	0.03 ***	0.03 ***	0.03 ***	0.03 ***	0.02[a]
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Level 2	Coefficient (standard error)				
Frequency-based rating: Semantic deviation (FS)					
FS intercept (γ_{01})	-	0.34 ***	0.33 ***	0.34 ***	-0.23[b]
		(0.04)	(0.04)	(0.04)	(0.34)
FS slope (γ_{11})	-	-	-	-	0.03[c]
					(0.02)
Frequency-based rating: Minor grammar error (FG)					
FG intercept (γ_{02})	-	-	0.04	-	
			(0.04)		
Collapsed category of paraphrase (CCP)					
CCP intercept (γ_{02})	-	-	-	0.05	
				(0.04)	
Random effects					
Between-scale intercept variance (μ_0)	0.025	0	0	0	0
Within-scale variance (r)	0.187	0.184	0.184	0.184	0.183
Model fit					
χ^2 (*df*)	-288.32 (4)	-275.07 (5)	-274.53 (6)	-274.17 (6)	-273.63 (6)
AIC	584.63	560.14	561.05	560.33	559.26
BIC	601.33	581.01	586.10	585.37	584.30
Model comparison	vs. Null	vs. C1	vs. C2	vs. C2	vs. C2
Likelihood-ratio (df)	11.24 (1)***	26.49 (1) ***	1.09 (1)	1.81 (1)	2.88 (1)
Proportion reduction (Between-scale)					
vs. Null	0%	100.00%	100.00%	100.00%	100.00%
vs. C1	-	100.00%	100.00%	100.00%	100.00%
vs. C2	-	-	0%	0%	0%
Proportion reduction (Within-scale)					
vs. Null	2.36%	3.65%	3.87%	4.01%	4.23%
vs. C1	-	1.32%	1.54%	1.69%	1.91%
vs. C2	-	-	0.23%	0.38%	0.60%

Notes: *p<0.05 **p<0.01 ***p<0.001; [a] *p* = 0.12, [b] *p* = 0.51, [c] *p* = 0.09

The comparison of the five MLM models identified Model C2 as the best-fitting model for the current data, which entered prompt length as a between-scale predictor and FS as a within-scale predictor to the Null model. The equation for the final model is as follows:

Level-1 Model (Within-scale): $Y_{ij} = \beta_{0j} + \beta_{1j}$ (Prompt Length) + r_{ij}

Level-2 Model (Between-scale): $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (FS) + μ_{0j}

$\beta_{1j} = \gamma_{10} + \gamma_{11}$ (FS) + μ_{1j}

In the within-scale model, Y_{ij} is the ith item's item discrimination in the jth scale, which is a function of the intercept (β_{0j}), prompt length effect (β_{1j}), and r_{ij}, the extent to which the item discrimination for item i in scale j that is not fully explained by prompt length. In the between-scale model, the two regression coefficients derived from the within-scale equation for each scale (i.e., β_0 and β_1) served as an outcome in two regression analysis, respectively. FS predicts the variation across the scales in the intercept β_{0j} and in the slopes β_{1j}, respectively in the two Level-2 equations.

To conclude, the results of Model C2 revealed that the prompt length significantly impacted item discrimination across the ten different ordinal scales/ scoring methods, but the magnitude was very small (RQ 5.2a), which invites to examine impacts of other prompt features. Also, scales that employed FS showed significantly higher item discrimination (RQ 5.2b).

8.3 Discussion and Conclusion

Motivated by the importance of adjusting item difficulty for category adequacy, this chapter examined the impact of sentence length, a widely-known contributor to item difficulty (Bley-Vroman & Chaudron, 1994; Campfield, 2017; Graham et al, 2010; Vinther, 2002; Yan et al., 2016) (RQ5.2). Results from univariate regression analyses indicated that prompt length significantly impacted item difficulty, in line with previous studies. Interestingly, the number of syllables of a prompt, rather than word count, was found to significantly impact item difficulty, unlike in Campfield (2017). One potentially novel examination of the study is that the magnitudes of the positive relationships were comparable cross the scales and/or scoring methods, whether frequency-based or quality-based criteria were used, or whether the scales/rubrics were conducted with the large or small samples in the current study. The findings support controlling the number of syllables to adjust item difficulty, which is the current practice of the testing program, and the continued use for revised scales/scoring methods proposed in Chapter 7.

The total variation explained by sentence length, however, was not large. The large unexplained variation was due to the considerable variations among the items of similar length, particularly longer prompts, which is another important novel finding of the study about the EI context. Longer items showed larger variation of item difficulty than short(er) items, and larger variation of prompt length for lower item difficulty than for higher item difficulty. Qualitative examination of the prompts of similar sentence length but different item difficulty revealed a range of lexical and syntactic sophistication among the items. This finding raises the awareness of the limited effectiveness of sentence length as a singular measure of prompt complexity and tool for controlling item difficulty. The large variation in linguistic features also calls for investigation of prompt linguistic features to understand and manipulate EI item difficulty, beyond the sentence length. The call will be addressed in the next chapter.

In addition to item difficulty, this chapter presented the results on item discrimination as a function of sentence length of EI prompts. Results from a series of MLM analysis indicated that prompt length was significantly associated with item discrimination across the different scales. However, variance in item discrimination explained by prompt length was very small, which invites investigation of the impact of other item characteristics, such as prompt linguistic features, on item discrimination. Almost all of the variance in item discrimination among different scales was due to the use of frequency-based semantic rating, but its interaction with prompt length was not significant.

Sentence length has been widely used to control EI item difficulty due to the support from previous research and practical benefit for item development (i.e., easy manipulation). The findings from the chapter confirms that the importance of using syllable-based sentence length while revealing its limitation. The results from the chapter also add new insights into understanding item discrimination in relation with sentence length, which is much less investigated topic. Because item discrimination is another important item measurement statistic, the findings are important for item quality control, while serving as instrumental information for item development when test/item developers apply prompts of different lengths to consider both item difficulty discrimination.

CHAPTER 9. IMPACT OF PROMPT LINGUISTIC FEATURES ON EI ITEM DIFFICULTY AND DISCRIMINATION

Responding to the call for looking into prompt complexity beyond sentence length to explain variation EI item parameters Chapter 8, this chapter examined 47 linguistic features of the EI prompts, including sentence length, in terms of their predictability of EI item difficulty and discrimination, as well as possible interaction effects. A series of random forest (RF) models were run to answer the following questions:

RQ 6.1 (a) Which linguistic features of EI prompts are important to predicting EI item difficulty across the category levels and different scales? (b) Which features interact with category levels of difficulty and/or scales/scoring methods ? (Section 9.1)

RQ 6.2 (a) Which linguistic features of EI prompts are important to predicting EI item discrimination across the different scales? (b) Which features interact with scales/scoring methods? (Section 9.2)

RF analyses revealed that lexical sophistication and phrasal-level syntactic complexity were more important than sentence length in predicting item difficulty and discrimination. Overall, interaction effects were marginal, particularly smaller in predicting discrimination than item difficulty. The largest interaction was found between item category level and sentence length in predicting item difficulty levels, in that sentence length is more important for a higher item category level.

9.1 Predicting EI Item Difficulty: Important Prompt Linguistic Features and Interactions Across Item Category Levels and Scales / Scoring Methods (RQ 6.1)

Important linguistic predictors of EI item difficulty (IRT-based parameter a), including their interaction effects, were examined in relation to three different category-level based difficulty as an outcome: item difficulty at all levels, the paraphrase level, and the minor error/deviation level. The following sections report the best-performing models and important linguistic features and interactions.

9.1.1 Important Prompt Linguistic Features for Item Difficulty (RQ 6.1a)

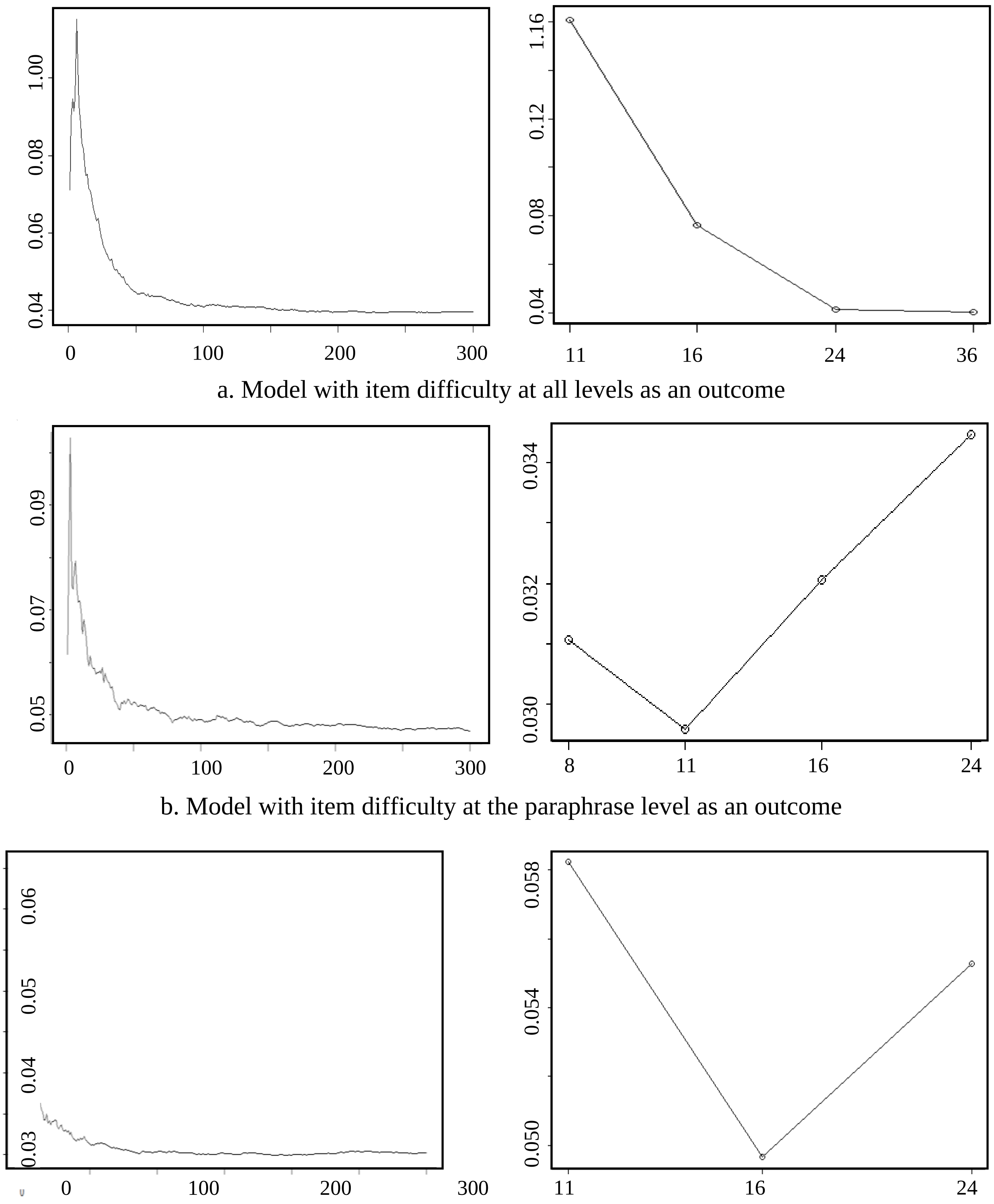
Model Building and Selection

To identify the best-performing model for item difficulty, a series of RF models were run. Table 9.1 provides the model specifications of nine representative models. First, three baseline models (i.e., RFb-B, RFbP-B, and RFbM-B) were run with the default values (mtry = 1/3 of the total number of variables, mtree = 300, node size = 5, randomly selected sample = 63.2% of the training sample). Using the baseline models, the number of trees (i.e., mtree) and random variables (i.e., mtry) for optimal modeling were identified, as illustrated in Figure 9.1. The large enough number of trees to get stably lowest OBB rates was 300 trees across the three outcome groups. The number of random variables (i.e., mtry) for the lowest out-of-bag (OBB) rate varied, 24, 11, and 16 for item difficulty at all, paraphrase, and minor error/deviation levels, respectively. Based on these mtry values, the three optimal RF models (i.e., RFb-O, RFbP-O, and RFbM-B) were built. For the item difficulty at the minor errors/deviation level, the baseline model, RFbM-B, was identified as the optimal model. The last three models, RGb, RGbP, and RGbM, implemented two more conditions (i.e., node size, sample fraction) for optimal modeling. Appendix 9.1 listed the sets of the conditions with the 10 lowest OBB error rates for each outcome group.

Table 9.1 Model Specification of the Thirteen RF Models

Model	Predictors	tree (#)	mtry	Node size	Sample fraction	Notes
Outcome of item difficulty at all levels (K = 49, N = 1440)						
RFb-B	item category levels, linguistic features, scales,	300	17	5	63.2%	baseline
RFb-O		300	24	5	63.2%	optimal mtry
RGb		300	35	3	80.0%	3 optimal conditions
Outcome of item difficulty at paraphrase (K = 48, N = 528)						
RFbP-B	linguistic features, scales	300	16	5	63.2%	baseline
RFbP-O		300	11	5	63.2%	optimal mtry
RGbP		300	31	3	80.0%	3 optimal conditions
Outcome of item difficulty at minor errors/deviation (K = 48, N = 475)						
RFbM-B	linguistic features, scales	300	16	5	63.2%	baseline, optimal mtry
RGbM		300	35	5	80.0%	3 optimal conditions

Notes: The *randomeForest* R package was used for the six RF models, and the *ranger* package for the three RG models.



a. Model with item difficulty at all levels as an outcome

b. Model with item difficulty at the paraphrase level as an outcome

c. Model with item difficulty at the minor error / deviation level as an outcome

Figure 9.1 The OBB Rates by the number of trees (left) and the number of random variables used RF models

Notes: x-axis – number of trees (left) or random variables (right), y-axis – OBB error rates

Based on the performance on the testing set, the best-performing model was selected for each outcome group. Table 9.2 listed the best-performing and/or selected models. See Appendix 9.2 for the performance of all 12 models. For the outcome group of item difficulty at the paraphrase level, Model RFbP-O was selected over the best-performing model, RGbP, given the minimal performance difference between the two models but higher generalizability and parsimoniousness of RFbP-O. RGbP used larger mtry (31) and sample fraction (80%) than RFbP-O (mtry = 11, sample fraction = 63.2%), which resulted in a more complex model with lower generalizability/randomness. The best-performing model was selected for the other groups.

Overall, the performance of the three selected models was (very) good. The correlations between the model-based predictions and empirical item difficulty values were high in that the prediction rates ranged from 0.93 (in RFbM-B, at the minor errors / deviation level) to 0.98 (in RFb-O, at all levels). The RMSE was (acceptably) low, which ranged from 0.16 of item difficulty (in RFbP-O, at the paraphrase level) to 0.22 (in RFbM-B, at the minor error / deviation level). Thus, the variable importance values of these models were obtained to answer RQ 6.1a.

Table 9.2 Performance of the Best-Performing and/or Selected Models

Model	Outcome	Mtry	Training set			Testing set	
	Item category level		Variance explained	RMSE (MSE)	Prediction (r)	RMSE (MSE)	Prediction (r)
RFb-O	all levels	24	97.81%	0.1987 (0.0395)	0.9923	0.1824 (0.0333)	0.9824
RFbP-O	paraphrase	11	95.29%	0.1627 (0.0265)	0.9747	0.1566 (0.0245)	0.9616
RGbP*	paraphrase	31	95.83%	0.1533 (0.0235)	0.9848	0.1458 (0.0212)	0.9654
RFbM-B	minor errors / deviation	16	90.85%	0.2165 (0.0469)	0.9708	0.2239 (0.0501)	0.9316

Note: *RGbP – best performing but not selected.

Important Prompt Features for Item Difficulty

The selected model for each outcome group—difficulty at all levels, the paraphrase level, and the minor error / deviation, henceforth, the *All-Levels*, *Paraphrase*, and *Minor D/E* models, respectively—estimated variable importance (VI) based on the increase in OBB MSE (Z score) when each predictor was replaced by random noise. The VI values of all predictors are provided in Appendix 9.3. Appendix 9.3 also presents increase in node purity as additional information, which indicates the effectiveness of split by each variable. In both increase in MSE and in node purity, higher values indicate greater importance. Results regarding node purity are not included in this section because node purity values are biased in general. Thus, VI in this chapter refers to the increase in MSE (Z score) only.

The predictors of the three models consist of linguistic and non-linguistic variables. While important prompt linguistic features are of the greatest interest of this section, the non-linguistic, test-structure related features are also worth paying attention to. This is because the performance provides the fit between the actual and model data structures and information for application of prompt linguistic features to adjusting item difficulty depending on item category levels and types of scales/ scoring methods.

Non-Linguistic Features The *scales / scoring methods* variable was a common non-linguistic predictor in the three RF models, and the *threshold levels* variable was used in the *All-Levels* RF model only. The *threshold levels* variable, the level of item category, was the most important predictor in the *All-Levels* model with VI of 231.82. Figure 9.2 illustrates the marginal effect of *threshold levels* on item difficulty in the *All-Level* model, which indicates the item difficulty is an ascending order from minor error/deviation to exact repetition as the test was intended. The highest importance of the *threshold levels* demonstrates that the RF model building considered the data structure of different groups of item difficulty by item category levels as expected and required.

The other non-linguistic predictor, *scales/ scoring methods*, was most important in the *Minor E/D* model (VI = 32.39) and second most in the *All-Levels* model (VI = 13.79), but least important at the paraphrase level, which was not recommended for inclusion in the *Paraphrase* model The marginal effects of the *scales/ scoring methods* are found in Appendix 9.4. The low VI in the *Paraphrase* model was expected because modifications of scales and scoring methods involved the minor and major errors / deviations categories. Thus, the importance (or non-importance) captured in the three models demonstrates that the models reflected the data structure.

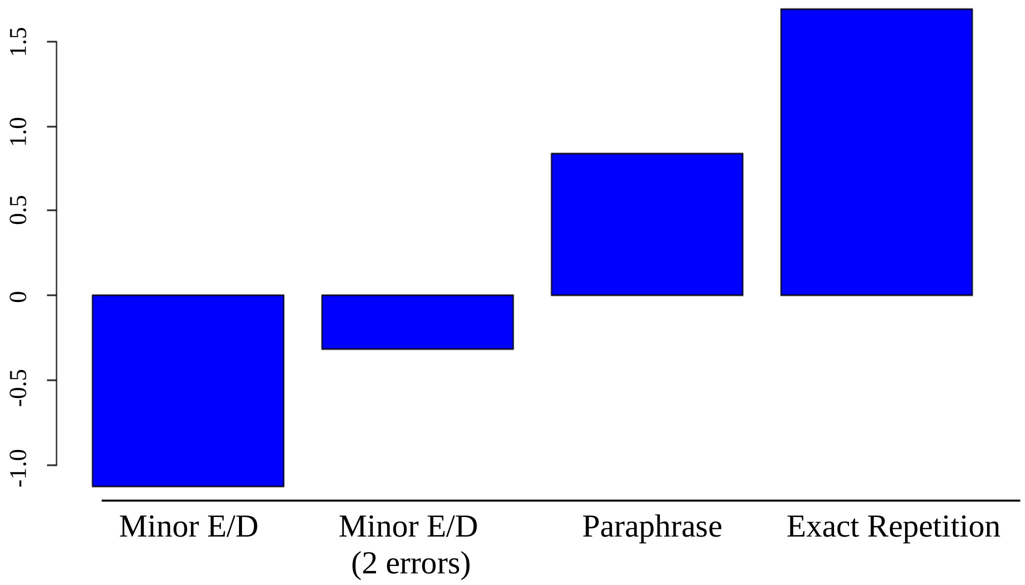
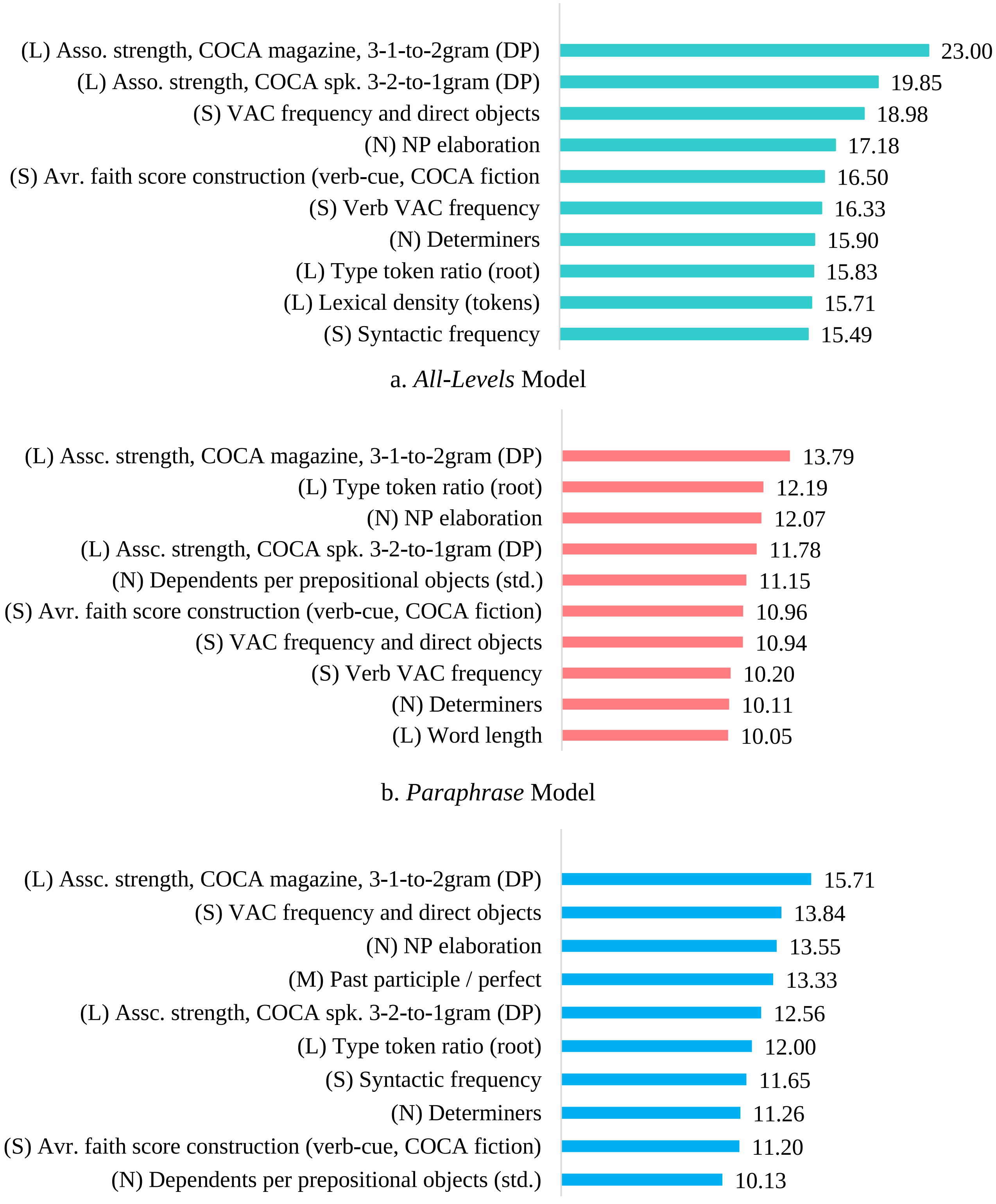


Figure 9.2 Marginal Effect of *Threshold Level* in the *all-level* Model

Notes: x-axis – threshold level, y-axis – item difficulty

Prompt Linguistic Features Among the 47 prompt linguistic features, Figure 9.3 plots the ten most important predictors of item difficulty in each model. The non-linguistic features were not included in the plots for simplicity. Note that, however, *scales*/*scoring methods* was more important than the linguistic features in the *All-Level* and *Minor E/D* models, and *threshold levels* in the *All-Level* model. Table 9.3 compares VI values and ranks by variable, which Figure 9.3 graphically provides in three separate plots. Table 9.3 further adds the linguistic types of the important linguistic features as well as all VI values and ranks of the predictors that ranked top ten in only one or two of the three models (i.e., *verb VAC frequency*, *past participle/perfect*, *dependents per prepositional object (std.)*, *word length*).



a. *All-Levels* Model

b. *Paraphrase* Model

c. *Minor E/D* Model

Figure 9.3 Top Ten Most Important Linguistic Features for Item Difficulty (VI Z-Score)

Table 9.3 Ten Most Important Linguistic Predictors for Item Difficulty* at the Paraphrase, Minor Error/Deviation, and All Levels

Category	Prompt linguistic features	Rank in Linguistic Variables (All Variables)			Variable importance		
		All	Paraphrase	Minor E/D	All	Paraphrase	Minor E/D
(L) Lexis: *n*-gram association strength	Association strength: trigram & unigram to bigram, COCA magazine (DP)	1 (3)	1	1 (2)	23.00	13.79	15.71
(L) Lexis: *n*-gram association strength	Association strength, trigram & bigram to unigram, COCA spoken (DP)	2 (4)	4	5 (6)	19.85	11.78	12.56
(S) Syntax: sophistication	VAC frequency and direct objects	3 (5)	7	2 (3)	18.98	10.94	13.84
(N) Noun phrase complexity	Noun phrase (NP) elaboration	4 (6)	3	3 (4)	17.18	12.07	13.55
(S) Syntax: sophistication	Average faith score construction (verb-cue, COCA fiction)	5 (7)	6	9 (10)	16.5	10.96	11.2
(S) Syntax: sophistication	Verb VAC frequency	6 (8)	8	14 (15)	16.33	10.20	9.19
(N) Noun phrase complexity	Determiners	7 (9)	9	8 (9)	15.90	10.11	11.26
(L) Lexis: variation	Type token ratio (root)	8 (10)	2	6 (7)	15.83	12.19	12.00
(L) Lexis: variation	Lexical density (tokens)	9 (11)	18	12 (13)	15.71	8.05	9.89
(S) Syntax: sophistication	Syntactic frequency	10 (12)	15	7 (8)	15.49	8.54	11.65
(D) Morphology	Past participle / perfect	12 (14)	20	4 (5)	14.31	7.45	13.33
(L) Lexis: difficulty	Word length	13 (15)	10	18 (19)	14.13	10.05	7.82
(N) Noun phrase complexity	Dependents per object of the preposition (std.)	18 (20)	5	10 (11)	11.11	11.15	10.13

Notes: *Variable importance was measured by increase in MSE (Z-score) when each predictor was replaced by random noise; NP – noun phrase; VAC - Verb-Argument Constructions

Overall, there were two main trends about the relationships between prompt linguistic features and EI item difficulty across the three models. First, sentence length (i.e., number of syllables) was not highly important for EI item difficulty compared to several lexical and syntactic features. Second, overall, lexical and syntactic measures were more important than morphological aspects. Also, lexical sophistication and phrase-level syntactic complexity were more important to predict item difficulty of the upper item category while overall or clausal-level syntax was more important for the lower level. The findings also bring two interesting points for discussion. First of all, fine-grained lexical and syntactic measures (e.g., association strength as a lexical measure, VAC-frequency based syntactic measures, and phrasal complexity) outperformed traditional (clausal) length-based or frequency-based measures across the models. Also, performance of some linguistic measures was notably different between the different item category levels, which suggests potential interaction effects.

- **Sentence length versus linguistic predictors**

Sentence length (i.e., number of syllables) was important to some degree only. Sentence length was the 11th most important in the *All-Level* model (VI = 15.24), and the 13th in the *Paraphrase* model (VI = 9.34), but less important in the *Minor E/D* model (VI = 6.35), ranking 22nd. The result is expected from Chapter 8, in that the number of syllables was significant but explained only a small portion of the variance. The (upper) mid ranks indicate that sentence length contributes to EI item difficulty to some degree, but the measure is not most sensitive or effective.

- **Lexical Sophistication, Diversity, & Difficulty (L)**

Among the three main linguistic components, lexical and syntactic features predicted item difficulty much more importantly than morphological predictors. Lexical items were most important in that the predictor with the largest VI was a lexical item across the three models, so was the second most important variable in the *Paraphrase* and *All-Levels* models.

The most important linguistic feature was a *n*-gram strength of association measure (i.e., the association strength: trigram and unigram to and bigrams derived from COCA Magazine using *delta* P). This measure represents how strongly trigrams or words cooccur with the following bigram. Prompts with high scores of this measure should have at least one or more bigrams in the first place that are common in the register of magazine writing, and trigrams would increase the scores. Among the current EI items, Item 8 scored zero (when rounded to the nearest hundredth), being the lowest, while Item 10 scored 0.46, being the highest (See below the two items). Item 10 includes several strongly associated pairs with a directional probability, such as *rank* and *second in*, *student* and *enrollment among,* or *foreign student enrollment* and *all public*. These *n*-gram pairs contributed to the high score of Item 10, which was one of the most difficult items. In contrast, Item 8 has some bigrams, such as *your lectures* and *class notes*, but the preceding word or trigrams, such as *record* or *if you record* did not seem to serve as a strong cue to elicit the bigrams in the written reference corpus.

Item 8 *If you record your lectures, you can revise your class notes later*. (Weak association)

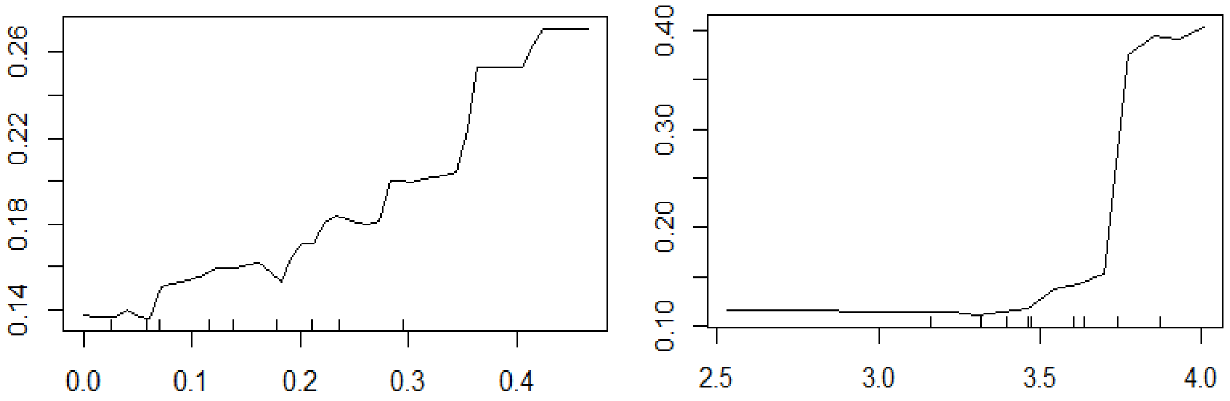
Item 10 *Purdue ranks second in foreign student enrollment among all public schools*. (Strong association)

The n-gram strength of association resulted in the greatest VI across the three models with fairly large differences from the second most important features. The VI ranged from 15.71 (in the *Minor E/D* model) to 23.00 (in the *All-Levels* model), which means on average, the absolute MSE values of individual tress increased as 15.71 to 23.00 times large as its standard deviation when the lexical variable was replaced by random noise. The marginal effects of the *n*-gram strength of association on the item difficulty are illustrated on the three partial plots of Figure 9.4 (see the plots on the left). Across all three models, the items with the stronger associations were more difficult in general, although (the number of) pivotal points varied between the models. Thus, strongly associated (tri-gram or longer) formulaic language in written discourse can be considered the most important and potentially most effective measure of EI item difficulty across the item category levels.

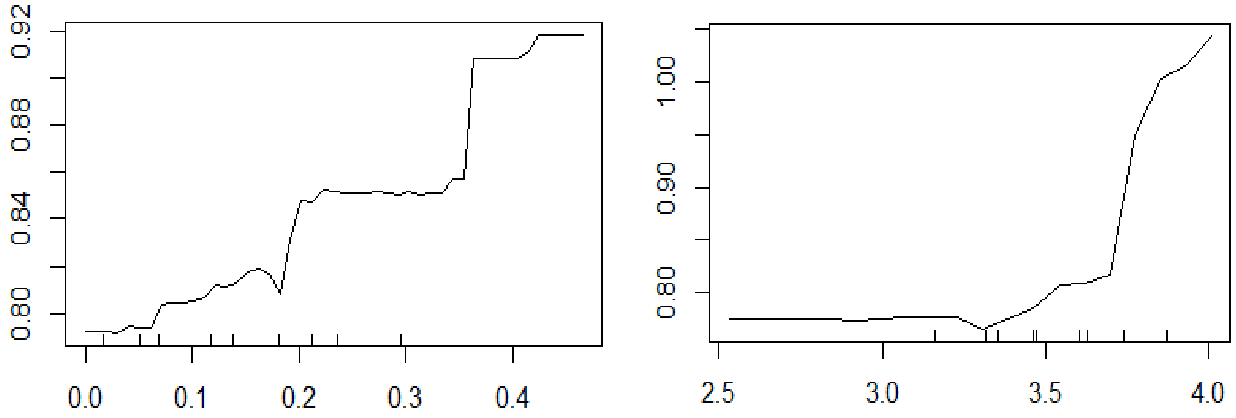
Interestingly, lexical measures seemed to be slightly more important for the item difficulty of the higher category than the lower category. The *Paraphrase* model had three lexical measures within the top ten. *Root TTR* (VI = 12.19), COCA Spoken-based *n*-gram association strength measure (VI = 11.78), and *Word length* (VI = 10.05) ranked second, fourth, and tenth. On the other hand, the spoken corpus-derived *n*-gram association strength (VI = 12.56) and root TTR (VI =

12.00) ranked lower in the *Minor E/D* model, and *word length* (VI = 7.82) was not even included the top ten measures.

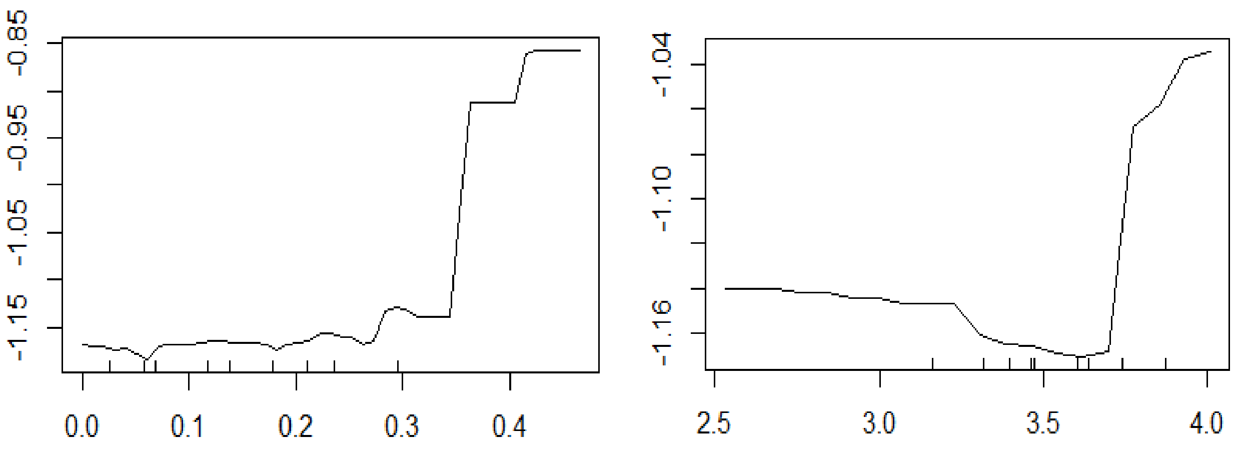
It is interesting that *root TTR* behaved differently in the *Minor E/D* model. The three plots in Figure 9.4 (see on the right) indicated that the relationship between root TTR and item difficulty was somewhat negative before the spike with approximately the TTR value of 3.75 while consistent and then somewhat upward in the other two models. Thus, root TTR seems most useful for prompts with reasonably high values unlike the association. Lexical density, the ratio of token-based content words, were much less important than root TTR, which includes both content and function words in both level-specific models but similar in the *All-Level* model. Word length was also less important than root TTR or lexical density.



All-Levels Model



Paraphrase Model



Minor E/D Model

Figure 9.4 Marginal Effects of Two Important Lexical Features: *Association Strength of Trigram and Unigram to Bigram, COCA Magazine (delta P)* (Left) and *Root Type-Token Ratio* (Right)

- **Syntactic Sophistication (S) and Noun Phrase Complexity (N)**

Syntactic features outnumbered lexical features in that six out of the ten most important linguistic features were syntactic measures in all models. However, clausal-length based, traditional measures did not rank high. All syntactic features within the top ten were fine-grained syntactic indices. One crucial reason was that 13 of the 16 clausal-length based measures were excluded from modeling because of their high correlation with each other, the component score, and sentence length. In this sense, sentence length might represent the clausal syntactic complexity measures. Sentence length and three clause-based measures (i.e., *verb phrases per T-unit*, *complex nominals per clause,* and *clauses per T-unit*) were mid-ranked, and the component score ranked even lower. Overall, sentence length ranked similar (in the *Minor E/D* model) or higher (in the *Paraphrase* and *All-Levels* models).

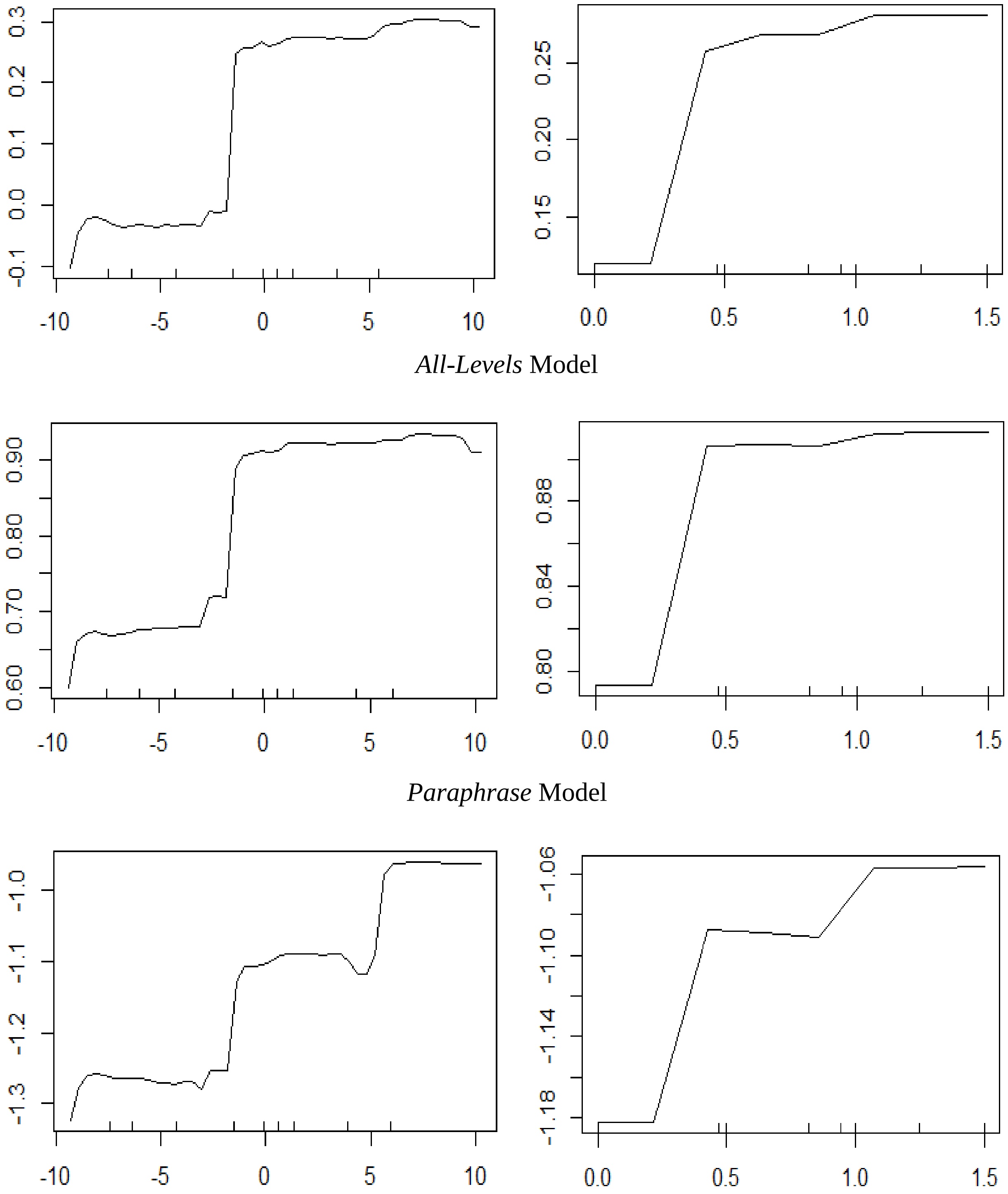
On the other hand, fine-grained syntactic measures were important across the item category levels. Notably, noun-phrase complexity was found to be as important as clausal-level syntactic sophistication in general. The importance particularly stood out for the paraphrase category. *NP elaboration*, the component measure of noun-phrase complexity indices, was the most important syntactic feature in the *Paraphrase* model, which ranked third (VI = 12.07), following the two lexical measures. At the low and all category levels, *NP elaboration* was the second most important syntactic feature, which ranked third in *Minor E/D* models (VI = 13.55) and fourth in the *All-*

Levels model (VI = 17.18). Figure 9.5 presents the positive relationship between *NP elaboration* and item difficulty in each model (see the three partial dependence plots on the left). The increasing patterns were consistent with all three models although the *Minor E/D* model has two major points where item difficulty changes drastically instead of one.

NP elaboration represents the structural complexity of the noun (and prepositional) phrases and nominals beyond simply counting the number (although related). Given heavier cognitive processing loads involved in prompts with more elaborate noun phrases and nominal, the positive relationships are reasonable. Examples from the current items clarify the relationship. Item 8 and Item 18 showed the lowest *NP elaboration* (-8.60, -9.31) and the values of Item 13 and Item 33 were the highest (8.58, 10.21). The structures of the noun phrases and nominals in Item 8 (i.e., *your lectures, your class notes*) and Item 18 (i.e., *I only have morning classes, this semester*) are not complex with only few prepositions and modifiers. In contrast, the noun phrases and nominals in Item 13 (i.e., *The amount of work involved in studying for final exams*) and Item 33 (i.e., his own story of finding an apartment) include complex structures by using diverse prepositions, prepositional objects with multiple dependents, and modifiers.

Item 8. *If you record your lectures, you can revise your class notes later*. (Low NP elaboration)
Item 13. *The amount of work involved in studying for final exams can overwhelm you*. (High NP elaboration)
Item 18. *It looks like I only have morning classes this semester*. (Low NP elaboration).
Item 33. *The senior student was talking about his own story of finding an apartment*. (High NP elaboration)

Note that, in the *Paraphrase* model, two other specific noun complexity measures, *Dependents per prepositional object* (VI = 11.15) and *Determiners* (VI = 10.11), ranked fifth and sixth, slightly higher than clausal syntactic measures. In the other two models, however, *dependents per prepositional object* ranked lower, tenth at most (in the *Minor E/D* model). Interestingly, from the three plots in Figure 9.5 (on the right), its marginal effect does not seem to form a drastically different pattern from that of *NP elaboration*: an increasing relationship in general, and one pivotal point at the paraphrase and all levels, and two for the lower category. Thus, the differences in VI ranks might be due to the performance of clausal syntactic features.



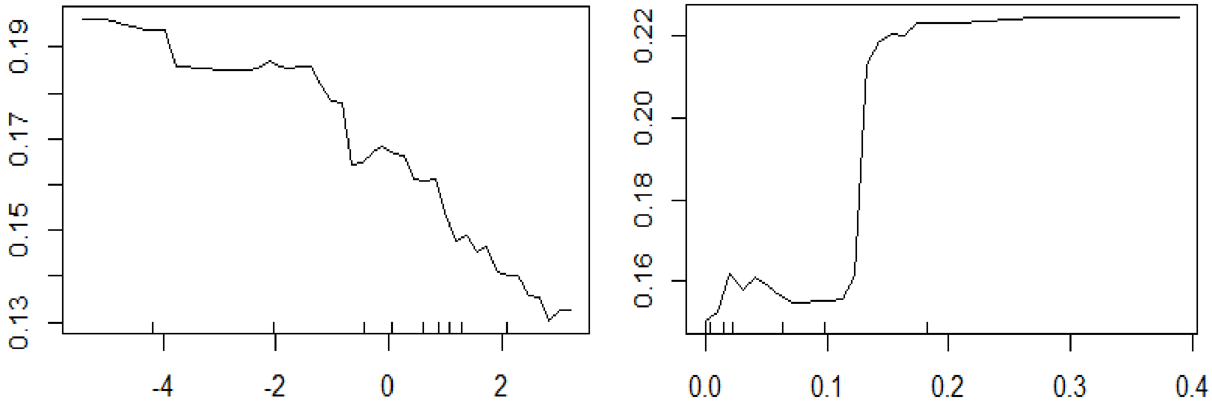
All-Levels Model

Paraphrase Model

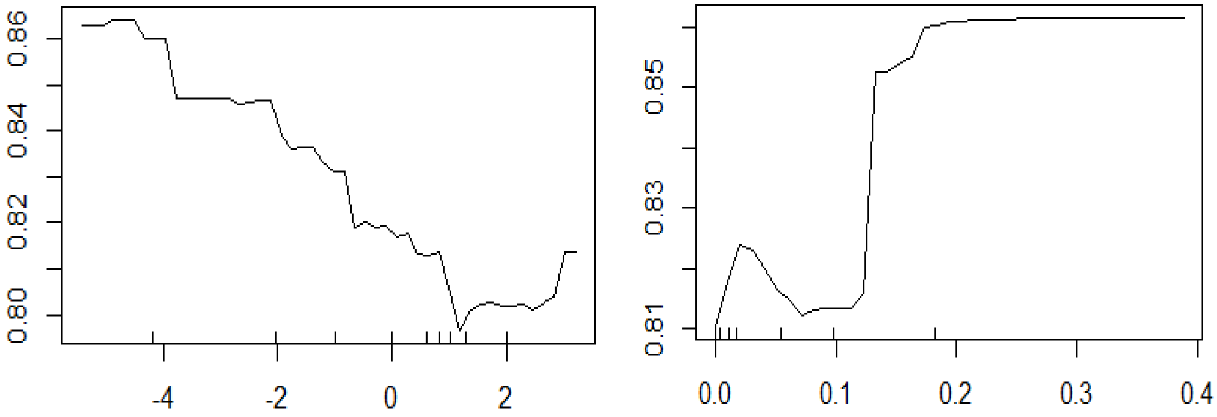
Minor E/D Model

Figure 9.5 Marginal Effects of Two Important Measures of Noun Phrase Complexity: *NP Elaboration* (Left) and *Dependents per Prepositional Object (std.)* (Right)

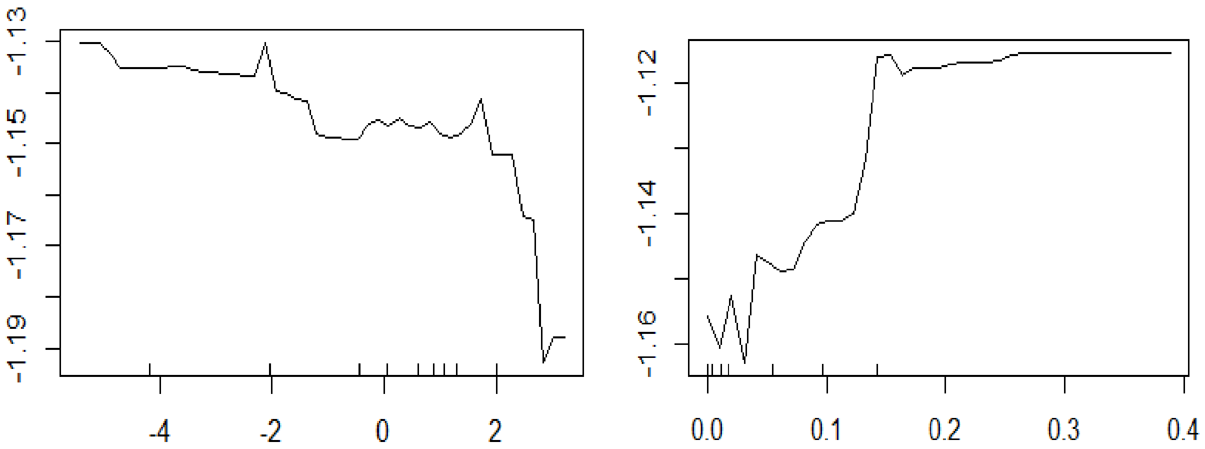
Unlike in the *Paraphrase* model, in the *Minor E/D*, clausal sophistication features ranked (somewhat) higher than phrasal complexity. *VAC frequency and direct objects* (VI = 13.84) was the most prominent syntactic feature, followed by a more comprehensive syntactic measure, *Syntactic frequency* (VI = 11.65), and a verb-cued association strength feature, *Average faith score construction (verb-cue, COCA fiction)* (VI = 11.20). In the *Paraphrase* model, however, a specific and association strength-based measure, *Average faith score of verb-cue construction* (VI = 10.96), was as important as *VAC frequency and direct objects* (VI = 10.94), which indicate how frequent verb-argument constructions (i.e., a verb and all its required arguments) appear in the reference corpus (i.e., COCA). *Syntactic frequency* (VI = 8.54), the most comprehensive, rough syntactic measure ranked lower. The *All-levels* model embraced the tendencies of both models. The mid-level comprehensive measure, *VAC frequency and directs objects* (VI = 18.98), still ranked highest among all syntactic measures, including noun complexity, but the verb-cued association measure ranked higher (VI = 16.50) than *Syntactic frequency* (VI = 15.49).



All-Levels Model



Paraphrase Model



Minor E/D Model

Figure 9.6 Marginal Effects of Two Important Measures of Syntactic Sophistication: *Syntactic Frequency* (Left) and *Average Faith Score Construction (Verb-Cue, COCA Fiction)* (Right)

Similar to the patterns in noun complexity, however, the overall tendencies in the marginal effects of syntactic sophistication measures did not considerably differ among the models. In Figure 9.6 above, overall, as the faith scores increased and *Syntactic frequency* decreased, item difficulty increased although the pivotal values of item difficulty increase/decrease and the slopes varied. The directions of relationships make sense, when examining the current EI prompts (See below). Item 3 obtained the lowest faith score and Item 25 was one of the highly scored items, both of which include the copula *be*. However, the probability of using the construction of *it's helpful to ask questions* (i.e., subject + verb + adjective complement + to infinitive) when a be verb *is* used compared with conditions not using *is* (Item 25) is much higher than the conditional probability of the construction, *which is the third largest city* (i.e., wh-subject + verb + noun complement), cued by *be* (Item 3). Verb-cued constructions with higher probability (i.e., higher faith scores) tend to be more formal, which makes processing more difficult. Likewise, frequently occurring constructions and lemmas tend to be easier. Item 30 and Item 40 are examples of items with the highest and lowest frequency. The less frequent lexico-syntactic structures in Item 40 elevates item difficulty.

Item 3 *Last month we travel**ed** to Chicago, which is the third largest city in the country*.
(Low faith score)

Item 25 *Sometimes, it’s helpful to ask questions in class as opposed to keeping them to yourself*. (High faith score)

Item 30. *This university has the third largest campus in the state*. (High syntactic frequency)

Item 40 *In other words, you must submit all your homework assignments on the course website*. (Low syntactic frequency)

- **Morphological Complexity (M)**

Morphological features of prompts (alone) were generally less important for the accuracy of prediction of item difficulty. There was only one morphological feature, *Past participle/perfect*, among the ten most important features, which ranked fourth in the *Minor E/D* model (VI = 13.33). *Past participle/perfect*, although being the most important morphological aspect across the models, was less important in the *All-Levels* (with ranking 14th, VI = 14.31) or the *Paraphrase* model (with ranking 20th, VI = 7.45). The difference suggests that the complexity escalated by the use of past participles and past perfect more greatly impacts the lower category, while the impact at the higher category might be embedded into or interact with other linguistic features.

In contrast to *Past participle/perfect*, *Plurals* was more important to the higher level. *Plurals* ranked12th (VI = 10.00) at the paraphrase level, but 21st (VI = 10.77) in the *All-Level* model and 30th (VI = 4.49) in the *Minor E/D* model. The importance of the other verb-related bound morphological aspects ranged moderate to low. Third-person singular for common verbs ranked 24th to 27th, being consistent across the models. Interestingly, third person copular *be* and past tense demonstrated mid-level importance for the paraphrase and all levels, but almost the lowest in the Minor E/D, ranking 44th and 45th out of 47 linguistic features. Note that the lower importance does not mean that the accuracy of using copular *be* or past tense is higher among the lower ability levels. Rather, the lower importance means that these features did not effectively distinguish the lower level because the performance of these bound-morphological aspects varied less at the lower level than at the higher level, and the variation was smaller compared with variations in other linguistic features.

9.1.2 Important Interactions Among Prompt Features for Item Difficulty (RQ 6.1b)

Model Building and Performance

To identify the pair-wise interaction effects, three SRC models (SRCb, SRCbP, SRCbM) were run with the three groups of outcomes, using the same conditions of the best-performing RF models. Table 9.4 shows the model specifications. Overall, the performance of the three models was not drastically different from the best-performing models with some variations among the three models. Model SRCbP, a model at the paraphrase level, performed slightly better than its corresponding RF model (MSE = 0.0233, RMSE = 0.1527, predictability = 0.9569), while Model SRCb, a model at all category levels, performed slightly worse than its RF counterpart (MSE = 0.0462, RMSE = 0.2150, predictability = 0.9751). The performance of the model at the minor error / deviation level, SRCbM, was worsened most (MSE = 0.0875, RMSE = 0.2958, predictability = 0.8409). The inaccuracy of approximately 0.3 is not ideal but not unacceptable considering the range of the item difficulty at the level. Thus, acknowledging the limitation, interaction analysis was continued.

Table 9.4 Model Specification of the Three SRC Models

Model	Item category level	Predictors other than linguistic features	tree (#)	mtry	Node size	Sample fraction
SRCb	All	threshold levels, scales /scoring methods	300	24	5	63.2%
SRCbP	Paraphrase	scales /scoring methods	300	11	5	63.2%
SRCbM	Minor E/D	scales /scoring methods	300	16	5	63.2%

Important Interaction

All possible pair-wise interactions of the variables were examined using the trees grown from the three SRC models. The interaction effect was assessed based on the difference between additive variable importance (i.e., the sum of the individual variable importance of two variables) and paired variable importance (i.e., variable importance of a paired variable). Overall, interaction effects were marginal across the models. The majority of the pairs showed a difference of 0.01 or smaller. Table 9.5 listed the pairs with the top five largest interactions, which are the pairs with the difference of 0.02 or larger (absolute values) in each model.

Table 9.5 The Top Five Largest Interaction Effects Across the three SRC Models

	Item	Variable Importance				
Variables (Interaction)	Category	Variable 1	Variable 2	Paired	Additive	Difference
Threshold level & Number of syllables	All	1.64	0.09	1.70	1.73	-0.03
Threshold level & Scales/scoring methods	All	1.64	0.03	1.65	1.67	-0.02
Threshold level & NP elaboration	All	1.64	0.07	1.69	1.71	-0.02
Number of syllables & Root TTR	All	0.09	0.06	0.17	0.15	0.02
Number of syllables & NP elaboration	Paraphrase	0.08	0.06	0.15	0.14	0.02*

Note: * The difference is not consistent because of rounding to the nearest hundredth.

The largest difference, 0.03, was detected between *Threshold level* (i.e., the category level at which item difficulty values) and *Number of syllables* (i.e., prompt length), followed by four other pairs: *Threshold*:*Scales/scoring methods*, *Threshold*:*NP elaboration*, *Number of syllables*:*Root TTR*, and *Number of syllables*:*NP elaboration*. No pairs from the *Paraphrase* or *Minor E/D* models showed a difference larger than 0.01 except for *Number of syllables* and *NP elaboration* in the *Paraphrase* model.

The difference in VI 0.03 can be considered marginal in general contexts. However, in the current context where the outcome variable is IRT-based item difficulty, even small amount of difference might indicate potentially meaningful interaction because the range of item difficulty is narrow in general. The VIs in the SRC models represent the absolute value of difference in OBB MSEs between trees with random permutation of a given variable and under observation of the variable while VIs in the RF models are normed values using standard deviation. Thus, the difference 0.03 in VI in SRC models means the maximum error of 0.17 in item difficulty—the value is RMSE, which can be obtained by taking the square root of 0.03. The difference of 0.17 in item difficulty is not large at all but not totally ignorable, either. Thus, the graphical descriptions were examined to evaluate the interaction effect.

Figure 9.7 graphically descries the relationship between item difficulty and sentence length by the three threshold levels (i.e., item category levels) based on raw values: exact repetition, paraphrase, and minor errors/deviations. Figure 9.8 presents the three partial dependence plots derived from the three RF models. According to Figure 9.7, across the three threshold levels, the patterns were fairly consistent, except for item difficulty of the items with 15 syllables. The overall pattern was increasing with a peak at 19 syllables, after which item difficulty went down. When the number of syllables were 15, the threshold of the exact repetition category slightly went up while the other two thresholds went down, slightly. These differences do not seem large enough to create an interaction effect.

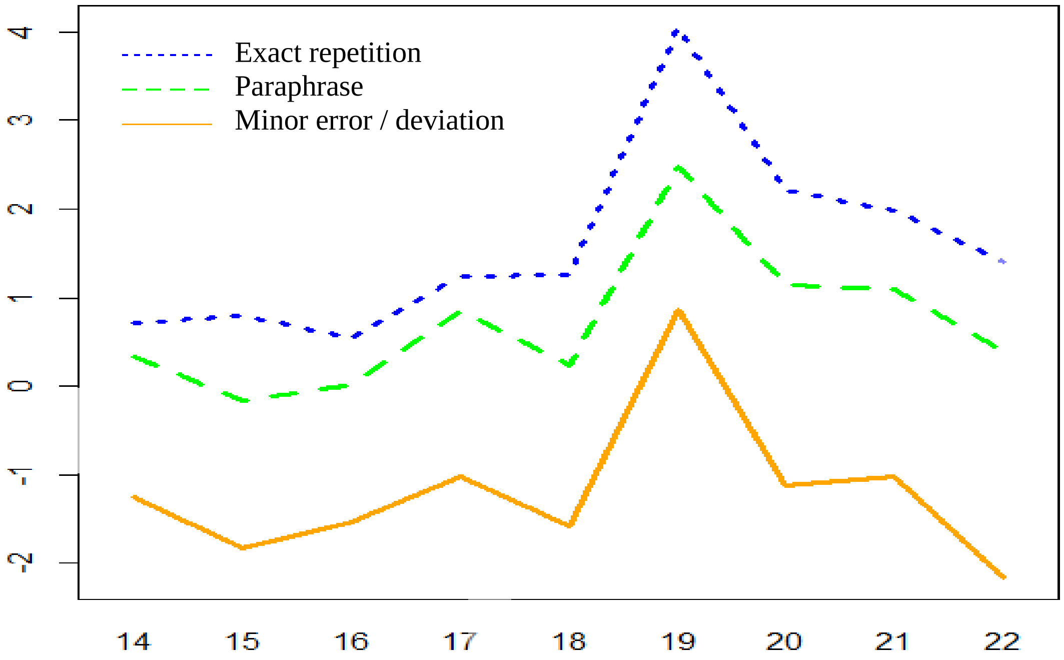
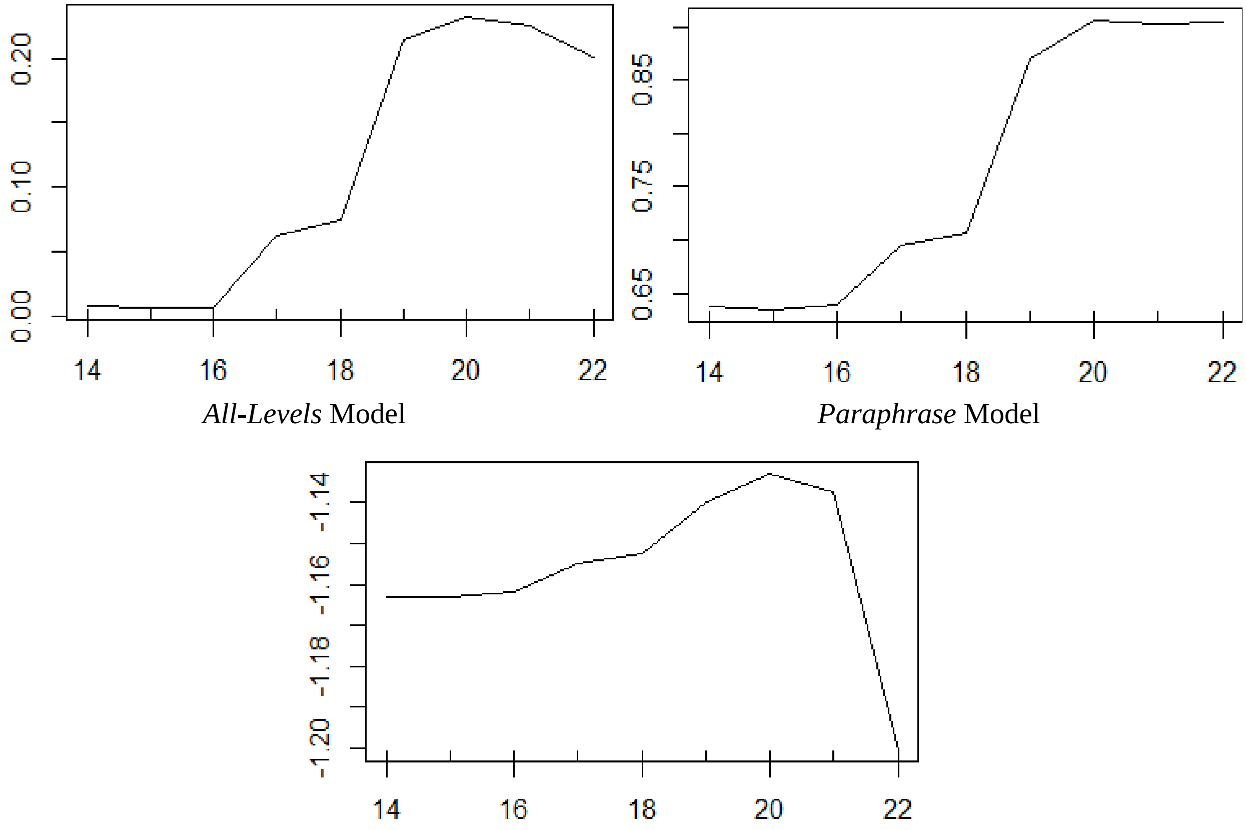


Figure 9.7 Item Difficulty by Sentence Length Between Three Threshold Levels

Notes: x-axis - sentence length (i.e., number of syllables); y-axis - item difficulty

The comparison of the marginal effects of sentence length described in Figure 9.8, however, clarifies the sources of the potential interaction between sentence length and threshold levels, focusing on two specific levels: the paraphrase and minor E/D levels. The magnitude of the association between item difficulty and sentence length was smaller in the minor E/D level although the relationship was all positive up until the peak at 20 syllables. Also, after the peak, the direction was changed differently. At the paraphrase level, item difficulty values were consistent among the items of 20 or more syllables while the relationship was changed to positive to negative in the minor E/R category. Because both paired variables do not have large univariate VI, the association will not affect modeling accuracy (Ishwaran, 2007), but the potential of interaction might be worth considering when an alternative dataset is used, for example, one that includes the paraphrase and all minor E/D levels only and exclude exact repetition and upper minor E/D levels.



All-Levels Model

Paraphrase Model

Minor E/D Model

Figure 9.8 Marginal Effects of Sentence Length (i.e., *Number of syllables*) on Item Difficulty in Three RF Models

In summary, the interaction effects were marginal at the *All-Levels* model, which means, in general, results about the linguistic predictors can apply across the item categories or scales and scoring methods while focusing on each individual linguistic feature. However, it is fairly probable that some linguistic features interact with category levels with more than marginal magnitude when only two levels, the paraphrase and minor E/D categories, are included. Thus, it be worth noting the difference in the important linguistic features found between the two levels in Section 9.1.1, that is the comparison of the *Paraphrase* and *Minor E/D* models, if any.

9.2 Predicting EI Item Discrimination: Important Linguistic Features and Interactions Across Scales / Scoring Methods (RQ 6.2)

In addition to EI item difficulty, important prompt features and interactions for item discrimination (IRT-based parameter a) were identified, using random forest modeling. The results are presented in the following sections, including the performance of the selected model.

9.2.1 Important Prompt Linguistic Features for Item Discrimination (RQ 6.2a)

Model Building and Selection

The best performing model was identified by running three RF models, following the same procedure taken when item difficulty was examined (See Section 9.1.1). The three models are: (1) a baseline RF model, RFa-B (mtree = 300, mtry = 17, node size = 5), (2) an optimal RF model using the *randomForest* package, RFa-O (mtree = 300, mtry = 24, node size = 5), and (3) an optimal RG model using the *ranger* package, RGa (mtree = 300, mtry = 35, node size = 5). Table 9.6 provides the model specifications and performance of the models. The baseline model, RFa-B performed best on the testing data. Although the accuracy and predictability of the model performance were not as high as those of the RF models for item difficulty, the error range of 0.2 can be considered acceptable for the outcome, item discrimination. Meanwhile, the smaller variance explained by the RFa-B model invites continued exploration for the sources of variance in EI item discrimination.

Table 9.6 Model Specification and Performance of the Three RF Models (N = 480, K = 48)

Model	Model specification				Performance (training set)				Performance (testing set)		
	mtree	mtry	node size	sample fraction	OBB MSE	OOB RMSE	variance explained	predictability	MSE	RMSE	predictability
RFa-B	300	17	5	63.2%	0.0470	0.2167	79.36%	0.9427	0.0392	0.1980	0.8132
RFa-O	300	24	5	63.2%	0.0466	0.2159	79.50%	0.9515	0.0403	0.2007	0.8095
RGa	300	35	5	70.0%	0.0705	0.2655	69.12%	0.8696	0.0833	0.2886	0.5805

Important Prompt Features for Item Discrimination

Based on the selected model (henceforth, the RFa model), VI values of all predictors (i.e., increase in MSE, Z score) were estimated. The predictors examined were 47 prompt linguistic features and one non-linguistic variable, *Scales/scoring methods*. The VI values of all predictors can be found in Appendix 9.5.

Non-Linguistic Features The types of scales and scoring methods (VI = 26.54) was the most important predictor to predicting item discrimination, and more influential than prompt linguistic features. The fact that Model RFa detected the systemic difference confirms that the model reflects the data structure as intended, as well as aligns well with the findings from Phase I.

Linguistic Features Figure 9.9 illustrates the top ten most important prompt linguistic features (which is the top 11 predictors if *Scales/scoring methods* is included). The general pattern of most important variables for item discrimination did not drastically differ from the tendency for item difficulty. Overall, lexical and syntactic features were more important than morphological features for item discrimination. Also, sentence length was far less important than linguistic features for item discrimination as well as than for item difficulty. Among the diverse lexical and syntactic measures, fine-grained measures outperformed than traditional measures. Lastly, large VI from scales and scoring methods increased the interest in its interaction with linguistic features in predicting item discrimination.

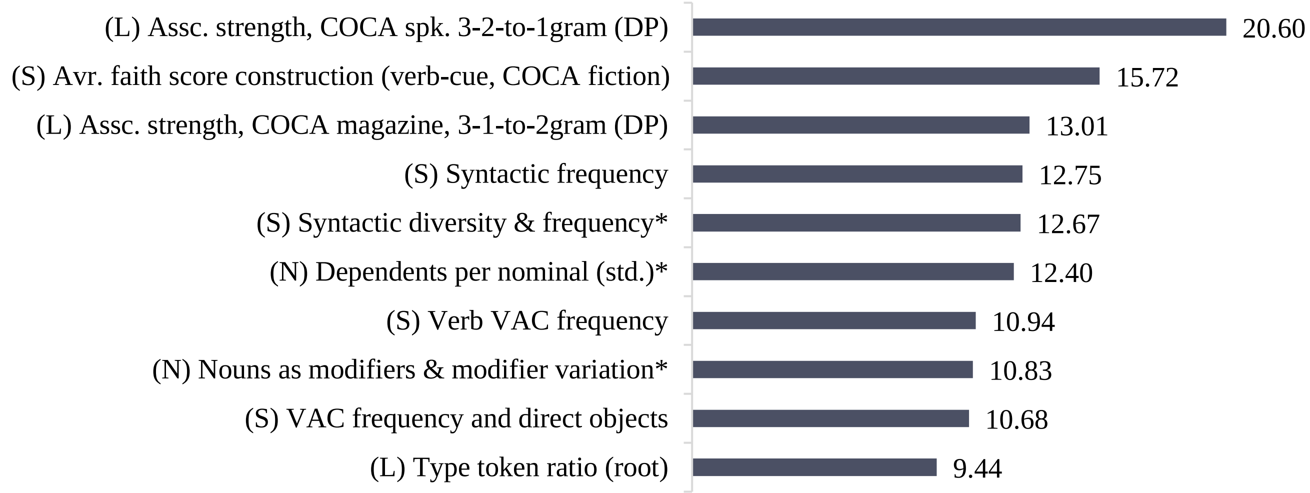


Figure 9.9 Top Ten Most Important Prompt Linguistic Features (and Scales/Scoring Method) for Item Discrimination (VI Z-Score)

Notes: *variables that ranked beyond the top ten most important linguistic features for item difficulty; (L) - lexical sophistication, diversity, and/or difficulty, (N) - noun phrase complexity, (S) - syntactic sophistication

- **Sentence length versus linguistic predictors**

Sentence length (i.e., number of syllables) did not greatly contribute to predicting item discrimination. *Number of syllables* (VI = 6.03) ranked 21st among the linguistic features. The rank was much lower than 11th in the *All-Level* model (VI = 15.24) and the 13th in the *Paraphrase* model (VI = 9.34), but similar to 22nd in the *Minor E/D* model (VI = 6.35) for item difficulty. Sentence length might serve as a rough, convenient measure, but it was not very sensitive to predict item parameters, and less sensitive for item discrimination.

- **Lexical sophistication, diversity, and/or difficulty (L)**

Consistent with the findings regarding item difficulty, lexical measures were very important for item discrimination. Two notable points were revealed. First, lexical measures ranked higher than syntactic features, particularly than clausal syntactic measures. Second, as for item difficulty at the paraphrase and all levels, fine-grained, *n*-gram association strength measures ranked higher than traditional, lexical diversity (e.g., TTR) or difficulty measures (e.g., word length). Two *n*-gram association strength measures, one derived from COCA Spoken (VI = 20.60) and the other on COCA Magazine (VI = 13.01), ranked first and third among all linguistic features, in order. Meanwhile, traditional measures, *root TTR* (for lexical diversity) and *world length* (for lexical difficulty) ranked tenth and twelfth.

- **Noun Phrase Complexity (N) and Syntactic Sophistication (S)**

Syntactic features were predominant in that noun phrase complexity and syntactic sophistication combined accounted for seven out of the ten most important linguistic features. Similarly for item difficulty, clausal-length based, traditional measures were found less important than fine-grained measures, none of which ranked within the top ten. However, one difference was that *the number of complex nominals per clause* ranked twelfth, higher than sentence length, unlike for item difficulty. Two noun phrase complexity measures ranked within top ten, *Dependents per nominal*, and *Nouns as modifiers and modifier variation*, further emphasize the degree of complexity of nominal structures, beyond simply counting the number of nominals for item discrimination.

Among the syntactic features, however, syntactic sophistication was more important than noun phrase complexity in general. There were five syntactic sophistication measures ranked within the top ten, which far outnumbered two noun phrase complexity measures. More interesting fining is that four out of the five syntactic sophistication measures were component measures, but most important one was faith scores of verb-cued constructions (VI = 15.72), which ranked second among all linguistic features. Verb-cued construction was also important for item difficulty at the paraphrase and all levels, but it was not as important as it ranked second for item discrimination. This syntactic association strength measure, along with the two COCA-derived lexical measures, reiterates the usage-based approach to understanding item discrimination.

- **Morphological Complexity (M)**

Morphological features were not as important as lexical or syntactic features for item discrimination. The most influential morphological feature was *Plurals* (VI = 7.29), which ranked 18th. The findings about morphological features should be carefully interpreted. The small VI values stress that simple counts of certain or overall morphological features present in the prompt are not sensitive independent measures. Despite the small VI values, it is reasonable to consider that diversifying morphological features contributes to item discrimination because morphological complexity is embedded into other important syntactic features, such as modifier variation, dependents of nominals, and syntactic frequency, for example.

9.2.2 Important Interactions Among Prompt Features for Item Discrimination (RQ 6.2b)

Model Building and Performance

Identification of potential interaction effects was pursued by running Model SRCa, using the best-performing RF model's specification (mtree = 300, mtry = 17, node size = 5). Table 9.7 shows the model specification and performance. With the training set, the model explained 71.38% of the variance in item discrimination, which was fairly less than the RFa model. The model performance was notably worsened with the testing set, which indicated that the model was overfit (MSE = 0.06, RMSE = 0.25, predictability = 0.68). Given the comprehensive coverage of prompt linguistic features in the SRCa model, sources other than linguistic features need to be explored to explain the variance.

Table 9.7 Specfication and Performance of Model SRCa

Model specification				Performance (training set)				Performance (testing set)		
mtree	mtry	node size	sample fraction	OBB MSE	OOB RMSE	variance explained	predictability	MSE	RMSE	predictability
300	17	5	63.2%	0.0670	0.2569	71.38%	0.8493	0.0634	0.2518	0.6760

Important Interaction

Based on the SRCa model, all possible pair-wise interactions between the 48 variables were examined. The results indicated that no pairs had difference of 0.01 or larger between the paired and additive VI values. The largest difference was almost zero, 0.0045, which was found in the pair of the two COCA-derived n-gram association strength measures. Thus, it was with strong confidence that interaction effects did not exist among the prompt linguistic features and scales/scoring methods in predicting item discrimination.

9.3 <u>Discussion and Conclusion</u>

Noting the centrality of prompt linguistic features to EI item difficulty as well as item development, Chapter 9 investigated important linguistic features for item difficulty. Particularly in relation to the exploration of optimal scales and scoring methods within the IRT framework in previous chapters, the non-linguistic factors were considered for modeling, as well. In addition, another important item parameter, item discrimination, was examined in search for important linguistic features. Results pointed out several interesting findings concerning the impacts of prompt linguistic features on item parameters, which makes implications not only for item development and scoring but also for the understanding of the L2 proficiency construct and its development trajectory, and for the directions of L2 English instructions for the current population (i.e., L2 college students of English)

9.3.1 Important Prompt Linguistic Features and Interactions for Item Difficulty

Concerning RQ 6.1, the RF analysis based on linguistic features combined with scales/scoring methods (and category levels in the *All-Level* model) predicted with high correlation (i.e., r = 0.93 to 0.98) and accuracy (i.e., RMSE = 0.18 to 0.22). Although sentence length contributed to some degree, most important linguistic variables were *n*-gram association strength (lexis) and fine-grained noun phrase complexity and syntactic sophistication. More interestingly, lexis, noun phrase complexity, and syntactic association strength were more important in the paraphrase category, while syntactic sophistication was more important for the lower category level. Meanwhile, throughout the models and pairs, interaction effects were very minimal, with some potential interaction between longer sentences and category levels only.

Overall, these findings confirm the previous research that found the important contribution of prompt lexical and syntactic linguistic features to EI item difficulty (Graham, McGhee, & Millard, 2010; Perkins et al., 1986; Ortega, 2000). The findings, however, provide further insights into specific types and characteristics of lexical and syntactic features that are important to EI item difficulty beyond traditional linguistic measures and sentence length. Aligning with previous research that found the importance of phrasal complexity (Biber et al., 2011; Kyle & Crossley, 2018), syntactic association strength and VAC frequency (Kyle, 2016), and lexical association strength (Kyle et al., 2018) on writing or speaking (Zhou, 2020), the current study offers the EI context as additional evidence. Interestingly, the outperformance of the fine-grained lexical and syntactic measures over classic measures for EI item difficulty in current study is in line with writing context more than speaking, although EI has often been used as a measure of oral proficiency. For example, Zhou's (2020) analysis noted that human rated ITA speaking test scores were more highly correlated with Lu's (2010) classic measures based on T-units or clauses than with VAC frequency or nominal complexity. The difference could be due to the different research design and methods. The current study examined IRT-based item difficulty extracted from human ratings rather than examinees' scores themselves, and the main analysis is RF regression, which is a non-linear, machine learning approach, rather than conventional statistical approach (e.g., regressions, MLM, correlational analysis). A more plausible explanation might be that the difference might have been led by the different construct of EI from that of free responses. As repeatedly corroborated in literature, EI can measure implicit knowledge (Bowles, 2011; Ellis, 2005; Erlam, 2006; Serafini, 2013) and/or processing competence (van Moere, 2012), which applies to speaking and writing as core aspects of language proficiency rather than being limited to speaking. Considering that formality and infrequency are representative characteristics of academic and/or written discourses, the findings seem reasonable. Future studies are invited to further clarify the construct of EI in relation to writing and speaking by examining the relationships between lexical and syntactic features and item difficulty focusing on tasks with different modality, including EI, on the same population.

Given the representation of EI item difficulty regarding L2 proficiency, the findings also inform us of the relationship between complexity and accuracy in the EI context. Because accurate repetition is central to EI scoring, which directly evaluates grammatical and semantic accuracy, the relationship between prompt linguistic features and EI item difficulty can be interpreted as one between complexity and accuracy. Information from fine-grained measures (e.g., lexical and syntactic association strength measures, corpus-referenced clause-level syntactic frequency, phrase-level syntactic complexity) supports understanding the relationships among the L2 proficiency components from a usage-based approach.

Another distinctive contribution from the current study is to address the predictability of item difficulty by prompt linguistic features between high and low category levels as well as across the levels. From the level-specific information, meaningful implications are made for L2 testing, proficiency research, and instruction. While the n-gram association measure ranked the highest for all levels, lexis and noun phrase complexity were more prominent at the higher category level than clausal level syntactic features. Thus, it is recommended that item development should consider enough variations among these features accordingly. This level-specific information can also be useful for scoring rubric, and potentially feature selection for automating scoring. Furthermore, the between-category level differences reveal L2 development trajectory for the target group, L2 English college students (and potentially L2 late teenagers and adults in general): (1) syntactic development from clausal to phrasal level, (2) later development of lexical sophistication than syntactic development, and (3) (although less important for item difficulty than lexical and syntactic features) delayed refinement of some morphological uses (e.g., plurals). The trajectory also provides guidance for L2 instruction. For example, for the examinees of the current test and similar L2 populations and contexts, L2 instructions are strongly encouraged to make sure to include formulaic languages (i.e., *n*-grams) with strong associations in general. Also, for upper-level students, instructions can focus more on sophisticated lexis, phrasal complexity, verb-specific clausal structures (i.e., VACs), and pluralization while low-level students might benefit from instruction that aim to broaden the syntactic range by including less frequently used structures and correctly use past participles and past perfects more immediately.

The predictability in relation to scales and scoring methods also offers useful information for the testing program and L2 testing research. Combining categories or modifying semantic or grammatical rating criteria impacted the item difficulty in general, but important linguistic features did not vary across the scales or scoring methods. The consistency indicates that consideration of linguistic features or applications of the implications aforementioned can be consistent regardless of the selection among the scales or scoring methods suggested in the current dissertation and similar revision approaches.

9.3.2 Important Prompt Linguistic Features and Interactions for Item Discrimination

Expanding on item difficulty, important linguistic features and interactions were examined for item discrimination (RQ 6.2). The examination addressed the lack of information on the relationships between prompt linguistic features and item discrimination in general as well as information on overall impact on item parameters (i.e., item difficulty and item discrimination) when item developers manipulate prompt linguistic features for item difficulty. No drastic distinction was made between the models for item difficulty and item discrimination. Similarly for the prediction of item difficulty at all levels, a n-gram association strength was most important. Noun phrase complexity was also important as well as clausal syntactic complexity, but clausal measures were slightly more prominent. Meanwhile, one subtle difference was noted: the largest VIs of association strength measures in both lexis and syntactic sophistication. The findings emphasize the importance of prompt linguistic features for item discrimination as well as item difficulty, while strongly recommending fine-grained measures, including noun-phrase complexity.

CHAPTER 10. CONCLUSION

This dissertation explored L2 English EI rating scales and scoring methods that measure grammatical and semantic accuracy in search for optimal measurement qualities, with special interests in item parameters, category adequacy, and misfit (in RQ1 through RQ4, Phase I). The project also identified important prompt linguistic features that predict item difficulty and discrimination of the EI items across different scales and scoring methods (in RQ5 and RQ6, Phase II).

In Phase I, answering RQ 1 (in Chapter 4) identified the 3-cateogry scale most optimal to measure grammatical and semantic accuracy in terms of adequacy of item parameter values, category functioning, and model, item and person fit, when using a series of GRM models. The 3-category scale collapsed the paraphrase into exact repetition category, and the lowest (i.e., omission, irrelevant, incomprehensible responses) into the major errors category. Particularly, appropriate item difficulty values of the paraphrase and exact repetition categories were correlated to the adequacy of category and scale usages. Expanding on the measurement qualities obtained by RQ1, answering RQ2 (in Chapter 5) found fewer misfitting persons with lower proficiency and higher frequency of unexpected responses in the lowest and highest categories, while qualitative examination for RQ3 (in Chapter 6) identified diverse sources of person and item misfit, among which two sources stood out: the inconsistency of distinguishing minor versus major semantic errors and the wide range of grammatical accuracy in the minor error category. The information led to suggestion and examination of revised scales / scoring methods / rubrics, which answered RQ4 (in Chapter 7). Among alternatives, the 4-category ordinal scale, a modification of the 3-category scale from RQ1, was found to be most optimal, which rated grammatical and semantic accuracy based on quantity rather than quality. The findings in Phase I highlighted the importance of scales and scoring methods to optimizing measurement qualities, particularly concerning specific categories, while provided specific information for future scale and rubric development and revision.

Moving on to Phase II, RQ5 (in Chapter 8) and RQ6 (in Chapter 9) investigated the impact of prompt linguistic features, including sentence length, on item difficulty and item discrimination across different scales (and ability levels). Results from univariate and multi-level modeling indicated that sentence length significantly explained item difficulty, aligned with previous studies, but the variance explained was not large. Further investigation using RF modeling revealed greater importance of corpus-based lexical measures and phrasal level syntactic complexity rather than conventional lexical diversity or clausal length-based syntactic complexity to predicting item difficulty, particularly for higher ability level. The findings highlighted the need to consider a usage-based approach to developing EI items and the differences across the ability levels.

The current dissertation has some limitations. The item bank is small, and the range of examinee proficiency is narrow, which limits the application of the findings, in general. In addition, the proposed scales and rubrics should be further examined with feasibility related to rater pools and contexts. Future studies are invited to address these issues.

REFERENCES

Akakura, M. (2012). Evaluating the effectiveness of explicit instruction on implicit and explicit L2 knowledge. *Language Teaching Research*, *16*(1), 9–37. https://doi.org/10.1177/1362168811423339

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355-377. https://doi.org/10.1177/0265532210364404

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *TESOL Quarterly*, *45*(1), 5-35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639-668. https://doi.org/10.1093/applin/amu059

Bley-Vroman, N., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In M. G. Tarone, & A. D. Cohen, *Research methodology in second-language acquisition* (pp. 245–261). Lawrence Erlbaum.

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition*, *33*, 247– 271. https://doi.org/10.1017/S0272263110000756

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley.

Bowden, H. (2016). Assessing second-language oral proficiency for research. *Studies in Second Language Acquisition*, *38*, 647–675. https://doi.org/10.1017/S0272263115000443

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345-370. https://doi.org/10.1007/bf02294361

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: User guide*. Scientific Software International.

Campfield, D. E. (2017). Lexical difficulty–using elicited imitation to study child L2. *Language Testing*, *34*(2), 197-221. https://doi.org/10.1177/0265532215623580

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chaudron, C., Prior, M., & Kozok, U. (2005). Elicited imitation as an oral proficiency measure. Paper presented at the 14th World Congress of Applied Linguistics, Madison, WI.

Christensen, C., Hendrickson, R., & Lonsdale, D. (2010). Principled construction of elicited imitation tests. Paper presented at the Language Resources and Evaluation Conference, Malta. http://www.lrec-onf.org/proceedings/lrec2010/pdf/672_Paper.pdf

Cook, K., McGhee, J., & Lonsdale, D. (2011). Elicited imitation as a prediction of OPI scores. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 30–37). Portland, OR: Association for Computational Linguistics.

Cox, T., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal*, *29*, 601–618. https://doi.org/10.11139/cj.29.4.601-618

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich.

Crossley, S. A., & Kyle, K. (2018). Analyzing spoken and written discourse: A role for natural language processing tools. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.) *The Palgrave handbook of applied linguistics research methodology* (pp. 567-594). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-59900-1_25

de Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Deshors, S. C. (2020). English as a Lingua Franca: A random forests approach to particle placement in multi-speaker interactions. *International Journal of Applied Linguistics*, *30*(2), 214-231. https://doi.org/10.1111/ijal.12275

Deshors, S. C., & Gries, S. T. (2020). Mandative subjunctive versus should in world Englishes: a new take on an old alternation. *Corpora*, *15*(2), 213-241. https://doi.org/10.3366/cor.2020.0195

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*(2) 141–172. https://doi.org/10.1017/s0272263105050096

Ellis, R. (2009a). Implicit and explicit learning, knowledge, and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 3–26). Multilingual Matters.

Ellis, R. (2009b). Measuring implicit and explicit knowledge of a second language. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 31–64). Multilingual Matters.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(2), 339- 368. https://doi.org/10.1017/s0272263106060141

Embretson, E. S., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*(3), 464–491. https://doi.org/10.1093/applin/aml001

Erlam, R. (2009). The elicited oral imitation test as a measure of implicit knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 65–93). Multilingual Matters.

Erlam, R., & Akakura, M. (2016). New developments in the use of elicited imitation. In A. Mackey, & E. Marsden (Eds.) *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages.* (pp. 105-123). Routledge.

Erlam, R., & Loewen, S. (2010). Implicit and explicit recasts in L2 oral French interaction. *The Canadian Modern Language Review* / La revue canadienne des langues vivantes, *66*(6), 877-905. https://doi.org/10.3138/cmlr.66.6.877

Faqeih, H. I. (2012). *The effectiveness of error correction during oral interaction: Experimental studies with English L2 learners in the United Kingdom and Saudi Arabia* [Unpublished doctoral dissertation]. University of York.

Fiori-Agoren, M. L. (2004). *The development of grammatical competence through synchonous computer mediated communication* [Unpublished doctoral dissertation]. Pennsylvania State University.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology*, *107*(1), 4–29. https://doi.org/10.1037/a0037289

Fouly, K., & Cziko, G. (1985). Determining the reliability, validity and scalability of the graduated dictation test. *Language Learning*, *35*(5), 555–566. https://doi.org/10.1111/j.1467-1770.1985.tb00361.x

Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of Verbal Learning and Verbal Behavior*, *2*(2), 121-135. https://doi.org/10.1016/s0022-5371(63)80076-6

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*(4), 287–291. https://doi.org/10.1093/elt/41.4.287

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The Elicited Imitation Task. *Language Learning*, *66*, 419–447. https://doi.org/10.1111/lang.12157

Gass, S. (2018). SLA elicitation tasks. In A. Phakiti, P. I. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 313–337). Palgrave Macmillan.

Gathercole, S. E., & Baddeley, A. D. (1993). Phonological working memory: A critical building block for reading development and vocabulary acquisition?. *European Journal of Psychology of Education*, *8*(3), 259-272. https://doi.org/10.1007/bf03174081

Graham, R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 1604–1610). Paris, France, European Language Resources Association.

Graham, C. R., McGhee, J., & Millard, B. (2010). The role of lexical choice in elicited imitation item difficulty. In M. Prior, Y. Watanabe, & S. Lee (Eds.), In *Selected Proceedings of the 2008 Second Language Research Forum* (pp. 57–72), Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, Document #2385

Gries, S. T. (2020). On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*, *16*(3), 617-647. https://doi.org/10.1515/cllt-2018-0078

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician, 63*(4), 308–319. https://doi.org/10.1198/tast.2009.08199

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking, & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57-78).Degnon Associates.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Hendrickson, R., Aitken, M., McGhee, J., & Johnson, A. (2010). What makes an item difficult? A syntactic, lexical, and morphological study of elicited imitation test items. In *Selected Proceedings of the 2008 Second Language Research Forum* (pp. 48-56). Somerville, MA: Cascadilla Proceedings Project.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning*, *33*(3), 315–332. https://doi.org/10.1111/j.1467-1770.1983.tb00544.x

Hood, L., & Lightbown, P. (1978). What children do when asked to “say what I say” – Does elicited imitation measure linguistic knowledge? *Allied Health and Behavioral Sciences,* 1, 195–219.

Hood, L., & Schieffelin, B. B. (1978). Elicited imitation in two cultural contexts. *Quarterly Newsletter of the Institute for Comparative Human Development*, *2*(1), 4–12.

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

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, *1*, 519-537. https://doi.org/10.1214/07-EJS039

Ishwaran, H., & Kogalur, U. B. (2014). Random forests for survival, regression and classification (RF-SRC), R package version 1.6. *http://CRAN. R-project. org/package= randomForestSRC*.

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, *64*(1), 215-238. https://doi.org/10.3138/cmlr.64.1.215

Kahng, J., & Otonya, M. (2021). What Elicited Imitation Can Show Us: Quantitative and Qualitative Analyses of Longitudinal Changes. *TESOL Quarterly*, *55*(1), 284-295. https://doi.org/10.1002/tesq.3020

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

Kim. J. (2012). *The optimal conditions for form-focused instruction: Method, target complexity, and types of knowledge* [Unpublished doctoral dissertation]. Georgetown University.

Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an elicited imitation test for SLA research. *The Modern Language Journal*, *100*, 655–673. https://doi.org/10.1111/modl.12346

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S-A. H., Gustafsson, J-E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science*, *18*(1), 146–154. https://doi.org/10.1111/desc.12202

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532221994052

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

Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement*, *41*(5), 388-400. https://doi.org/10.1177/0146621617692978

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication [Doctoral dissertation]. Georgia State University.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. Behavior Research, 50, 1010-1046. https://doi.org/10.3758/s13428-017-0924-4

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*(4), 451-475. https://doi.org/10.1177/0265532217713951

Lee, V. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, *35*(2), 125–141. https://doi.org/10.1207/S15326985EP3502_6

Li, S. (2010). *Corrective feedback in perspective: The interface between feedback type, proficiency, the choice of target structure, and learners' individual differences in working memory and language analytic ability* [Unpublished doctoral dissertation]. Michigan State University.

Li, X. (2020). *The Technical Qualities of the Elicited Imitation Subsection of The Assessment of College English, International (ACE-In)* [Unpublished doctoral dissertation]. Purdue University.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18–22. https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf

Logan, G. D. (1988). *Towards an instance theory of automatization*. *Psychological Review, 95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Lonsdale, D., & Christensen, C. (2011). Automating the scoring of elicited imitation. In *Proceedings of the ACL-HLT/ICML/ISCA Joint Symposium on Machine Learning in Speech and Language Processing*.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190-208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Maydeu-Olivares, A. (2015). Evaluating the fit of IRT modes. In S. P. Reise, & D. A. Revicki (Eds.) *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111-127). Routledge.

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713-732. https://doi.org/10.1007/s11336-005-1295-9

Markman, B. R., Spilka, I. V., & Tucker, G. R. (1975). The use of elicited imitation in search of an interim French grammar. *Language Learning*, *25*(1), 31–41. https://doi.org/10.1111/j.1467-1770.1975.tb00107.x

McDade, H. L., Simpson, M. A., & Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar: A question of validity. *Journal of Speech and Hearing Disorders*, *47*(1), 19-24. https://doi.org/10.1044/jshd.4701.19

Menyuk, P. (1971). *The acquisition and development of language*. Prentice Hall.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Miller, J. F. (1973). Sentence imitation in pre-school children. *Language and Speech, 16*(1), 1–14. https://doi.org/10.1177/002383097301600101

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2). 159-176. https://doi.org/10.1177/014662169201600206

Naiman, N. (1974). *Imitation, comprehension and production of certain syntactic forms by young children acquiring a second language* [Unpublished doctoral dissertation]. University of Toronto.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*(5), 625-632. https://link.springer.com/article/10.1007%2Fs10459-010-9222-y#citeas.

Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 2132–2137). Austin, TX: Cognitive Science Society.

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64. https://doi.org/10.1177/01466216000241003

Ortega, L. (2000). Understanding syntactic complexity: The measurement of change in the syntax instructed L2 Spanish learners [Unpublished doctoral dissertation]. University of Hawai'i.

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October). An investigation of elicited imitation tasks in crosslinguistic SLA research. Paper presented *at the Second Language Research Forum,* Toronto, Canada.

Ostini, R., Finkelman, M., & Nering, M. (2014). Selecting Among Polytomous IRT Models. In S. P. Reise, & D. A. Revicki (Eds.). *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 285 – 304). Routledge.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards, & R. W. Schmitt (Eds.), *Language and communication* (pp. 191–226). Longman.

Perkins, K., Brutten, S. R., & Angelis, P. J. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning*, *36*(2), 125–141. https://doi.org/10.1111/j.1467-1770.1986.tb00375.x

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology, 25*, 111-163. https://doi.org/10.2307/271063

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*(3), 595–626. https://doi.org/10.1111/lang.12010

Rebuschat, P., & Mackey, A. (2013). Prompted production. In C. A. Chappelle (Ed.), *The encyclopedia of applied linguistics*, vol. 5. Wiley-Blackwell.

Prutting, C. A., Gallagher, T. M., & Mulac, A. (1975). The expressive portion of the NSST compared to a spontaneous language sample. *Journal of Speech and Hearing Disorders, 40*(1), 40-48. https://doi.org/10.1044/jshd.4001.40

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4), 1–97. https://doi.org/10.1007/bf03372160

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*(2), 203–219. https://doi.org/10.1007/bf02291114

Sarandi, H. (2015). Reexamining elicited imitation as a measure of implicit grammatical knowledge and beyond …? *Language Testing*, *32*(4), 485–501. https://doi.org/10.1177/0265532214564504

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Serafini, E. (2013). *Cognitive and psychosocial factors in the long-term development of implicit and explicit second language knowledge in adult learners of Spanish at increasing proficiency* [Unpublished doctoral dissertation]. Georgetown University.

Shin, J-y. (2021). The use of stance in L2 first-year college writing. In M. Charles, & A. Frankenberg-Garcia (Eds.) *Corpora in ESP/EAP writing instruction: Preparation, exploitation, analysis* (pp. 123-146). Routledge.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*(4), 375-394. https://doi.org/10.1111/j.1745-3984.2005.00021.x

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*(1), 23-35. https://doi.org/10.1111/emip.12024

Sinharay, S., & Lu, Y. (2008). A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement*, *45*(1), 1-15. https://doi.org/10.1111/j.1745-3984.2007.00049.x

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*(3), 723–751. https://doi.org/10.1111/lang.12129

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323–348. https://doi.org/10.1037/a0016973

Sullivan, G. M., & Artino Jr., A. R. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*. *5*(4), 541-542. https://doi.org/10.4300/JGME-5-4-18

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson Education Inc.

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*(1), 120-151. https://doi.org/10.1177/0272431613511332

Tracy-Ventura, N., McManus, K., Norris, J., & Ortega, L. (2014) . ‘Repeat as much as you can’: Elicited Imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Proficiency assessment issues in SLA research: Measures and practices* (pp. 143-166). Multilingual Matters.

Trofimovich, P., & Baker, W. (2007). Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Applied Psycholinguistics*, *28*(2), 251-276. https://doi.org/10.1017/S0142716407070130

Underhill, N. (1987). *Testing Spoken Language: A Handbook of Oral Testing Techniques*. Cambridge University Press.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344. https://doi.org/10.1177/0265532211424478

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73. https://doi.org/10.1111/1473-4192.00024

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition*, *4*(3), 203–222. https://doi.org/10.1515/langcog-2012-0011

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software. 77. 1–17. https://doi.org/10.18637/jss.v077.i01

Wu, S-L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*(4), 680–704. https://doi.org/10.1111/flan.12063

Yan, X. (2015). *The processing of formulaic language on elicited imitation tasks by second language speakers* [Unpublished dissertation]. Purdue University.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497-528. https://doi.org/10.1177/0265532215594643

Zhou, Y. (2012). *Willingness to communicate in learning Mandarin as a foreign and heritage language* [Unpublished doctoral dissertation]. University of Hawaii at Manoa.

Zhou, Z. (2020). *Modeling statistics ITAs' speaking performances in a certification test* [Doctoral dissertation], Iowa State University.

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology*, *34*(4), 390-400. https://doi.org/10.1037/h0078865

APPENDIX

Appendix 3.1 List of Prompt Features Used as Predictors

No	Linguistic Component	Variable Abbreviated	Full/ Measurement
1	L	COCA_spoken_tri_2_DP (Assc. strength, COCA spk. 3-2-to-1gram (DP))	Association strength of tri- and bi- gram with unigram; reference corpus of COCA Spoken, Delta P
2	L	COCA_magazine_tri_DP (Assc. strength, COCA magazine, 3-1-to-2gram (DP))	Association strength of tri- and unigram and with bigram; reference corpus of COCA Magazine, Delta P
3	L	Word length	Average number of syllables per word
4	L	TTR_root	Type token ratio (root)
5	L	lexical_density_tokens	Number of content words tokens divided by number of total number of tokens
6	L	AWL	Academic word list (all)
7	L	mtld_original_aw	MTLD is based on the average number of tokens it takes to reach a given TTR value (.720)
8	S	Syntactic frequency	Factor score of: - percentage of constructions in text that are in reference corpus (all) - percentage of lemma construction combinations in text that are in reference corpus (all) - average construction frequency, log transformed (all) - average lemma frequency, log transformed (all)
9	S	Syntactic diversity & frequency	Factor score of: - construction type-token ratio (all) - main verb lemma type-token ratio (all) - lemma construction combination type-token ratio (all) average lemma construction frequency (types only) (all)

10	S	Verb VAC frequency	Factor score of: - average lemma construction combination frequency (all) - average lemma frequency (all) - average lemma construction combination frequency, log transformed (all) (standard deviation) - average lemma frequency, log transformed (all) (standard deviation) - average lemma construction combination frequency, log transformed (all) - nominal complements per clause - average lemma frequency (types only) (all) - adjective complements per clause
11	S	VAC frequency and direct objects	- average construction frequency (all) - average construction frequency (types only) (all) - average construction frequency, log transformed – all - direct objects per clause dependents per direct object
12	S	Avr. faith score construction (verb-cue, COCA fiction)	Average faith score verb (cue) - construction (outcome); reference corpus of COCA Spoken
13	S	C/T	Number of clauses per T-unit
14	S	CN/C	Complex nominals per clause
15	S	Pperfect	Number of past participles
16	S	prep_about	Number of prepositions *about*
17	S	prep_among	Number of prepositions *among*)
18	S	prep_at	Number of prepositions *at*
19	S	prep_by	Number of prepositions *by*
20	S	prep_during	Number of prepositions *during*
21	S	prep_for	Number of prepositions *for*
22	S	prep_in	Number of prepositions *in*
23	S	prep_of	Number of prepositions *of*
24	S	prep_on	Number of prepositions *on*
25	S	prep_to	Number of prepositions *to*
26	S	prepAll	Number of prepositions (all)
27	S	rcmod_dobj_deps_struct	Relative clause modifiers per direct object
28	S	VP/T	Verb phrases per T-unit

29	N	Nouns as modifiers & modifier variation	Factor score of: - nouns as a nominal dependent per nominal - nouns as a nominal dependent per nominal - dependents per direct object
30	N	NP (Noun phrase) elaboration	Factor score of: - prepositions per nominal - dependents per object of the preposition - prepositions per object of the preposition - prepositions per direct object - prepositions per nominal subject - adjectival modifiers per nominal dependents per nominal - dependents per nominal subject adjectival modifiers per nominal subject - adjectival modifiers per object of the preposition - adjectival modifiers per direct object determiners per nominal subject - passive nominal subjects per clause - dependents per direct object No pronouns) - dependents per object of the preposition No pronouns) - prepositions per clause - verbal modifiers per nominal - nominal subjects per clause - dependents per nominal complement
31	N	av_nominal_deps (Dependents per nominal (std.))	Number of dependents per nominal (std.)
32	N	ArticleA	Number of indefinite articles
33	N	ArticleThe	Number of definite articles
34	N	av_nsubj_deps_NN	Dependents per nominal subject (no pronouns)
35	N	deps_ prep_obj (std.)	Dependents per prepositional objects (std.)
36	N	Determiners	Number of determiners
37	N	pobj_NN_stdev	Dependents per object of the preposition No pronouns, standard deviation
38	N	prep_pobj_deps_NN_struct	Prepositions per object of the preposition (no pronouns)
39	M	ingGrnd	Number of gerunds
40	M	ingProg	Progressive tense

41	M	Past	Number of verbs with past tense
42	M	Plurals	Number of words with pluralization
43	M	possessives	Number of possessives
44	M	PPnPerfect	Number of past participles used for tense and modifiers
45	M	thirdSing	Number of verbs with third person singular -*s*
46	M	thirdSingBe	Number of third person singular be (i.e., *is*, *was*)
47	Length	NumSyl	Number of syllables

Notes: L - lexical measures; S - syntactic measures; M – morphological measures; Definitions from Kyle (2016) and Kyle, Crossley, & Berger (2018)

Appendix 4.1 Descriptive Statistics of Item Scores on the 5-Category EI Accuracy Scale

Item	Mean	SD	Median	Min.	Max.	Range	Skewness	Kurtosis	SE
Form 1 (N=193×12=2,280 observations)									
Q1	1.90	0.92	2	0	4	4	0.55	0.16	0.07
Q2	2.20	0.89	2	0	4	4	0.67	0.02	0.06
Q3	2.79	1.10	3	1	4	3	-0.18	-1.41	0.08
Q4	2.15	1.07	2	1	4	3	0.57	-0.93	0.08
Q5	2.28	1.12	2	0	4	4	0.04	-1.33	0.08
Q6	2.25	0.85	2	1	4	3	0.31	-0.50	0.06
Q7	2.66	1.10	2	1	4	3	0.09	-1.46	0.08
Q8	2.36	1.10	2	0	4	4	0.07	-1.18	0.08
Q9	1.94	0.72	2	1	4	3	0.42	0	0.05
Q10	1.10	0.74	1	0	4	4	0.74	1.17	0.05
Q11	1.48	0.70	1	0	4	4	0.74	1.88	0.05
Q12	2.12	0.90	2	0	4	4	0.59	-0.13	0.06
Form 2 (N=202×12=2,424 observations)									
Q13	1.34	0.90	1	0	4	4	0.78	1.18	0.06
Q14	2.47	0.95	2	0	4	4	-0.20	-0.22	0.07
Q15	2.25	0.89	2	1	4	3	0.55	-0.37	0.06
Q16	2.29	1.00	2	0	4	4	0.34	-0.60	0.07
Q17	1.95	1.04	2	0	4	4	0.85	-0.37	0.07
Q18	2.95	1.03	3	1	4	3	-0.34	-1.27	0.07
Q19	2.48	1.15	2	0	4	4	0.11	-1.36	0.08
Q20	2.56	1.14	2	1	4	3	0.09	-1.45	0.08
Q21	1.66	0.75	2	0	4	4	0.49	-0.11	0.05
Q22	1.86	0.79	2	0	4	4	0.61	0.38	0.06
Q23	1.60	0.75	2	0	4	4	0.65	0.79	0.05
Q24	1.98	0.95	2	0	4	4	0.32	0.01	0.07
Form 3 (N=204×12=2,448 observations)									
Q25	1.73	1.03	1	0	4	4	0.85	-0.02	0.07
Q26	1.62	0.94	1	0	4	4	0.81	0.26	0.07
Q27	2.23	0.86	2	0	4	4	0.24	0.40	0.06
Q28	2.00	0.75	2	0	4	4	0.14	-0.48	0.05
Q29	2.35	1.09	2	0	4	4	0.46	-1.05	0.08
Q30	2.54	1.27	2	1	4	3	0.07	-1.68	0.09
Q31	2.71	1.06	3	1	4	3	-0.04	-1.36	0.07
Q32	2.89	1.05	3	1	4	3	-0.18	-1.47	0.07

Q33	2.27	0.95	2	0	4	4	0.38	-0.62	0.07
Q34	1.90	0.76	2	0	4	4	0.58	0.74	0.05
Q35	2.35	0.97	2	0	4	4	0.43	-0.53	0.07
Q36	1.74	0.76	2	0	4	4	0.61	0.44	0.05
Form 4 (N=180×12=2,160 observations)									
Q37	1.93	0.90	2	0	4	4	0.79	0.11	0.07
Q38	2.26	0.96	2	1	4	3	0.56	-0.61	0.07
Q39	1.80	0.94	2	0	4	4	0.69	0.15	0.07
Q40	1.92	0.69	2	1	4	3	0.41	0.09	0.05
Q41	2.48	1.19	2	0	4	4	0.02	-1.20	0.09
Q42	2.23	1.01	2	1	4	3	0.56	-0.77	0.08
Q43	2.08	0.86	2	0	4	4	0.41	-0.32	0.06
Q44	2.77	1.03	2	1	4	3	0.02	-1.42	0.08
Q45	1.53	0.68	1	0	4	4	0.69	0.88	0.05
Q46	2.32	0.98	2	0	4	4	0.48	-0.63	0.07
Q47	2.02	0.77	2	0	4	4	0.49	0.40	0.06
Q48	2.58	0.82	2	1	4	3	0.24	-0.68	0.06

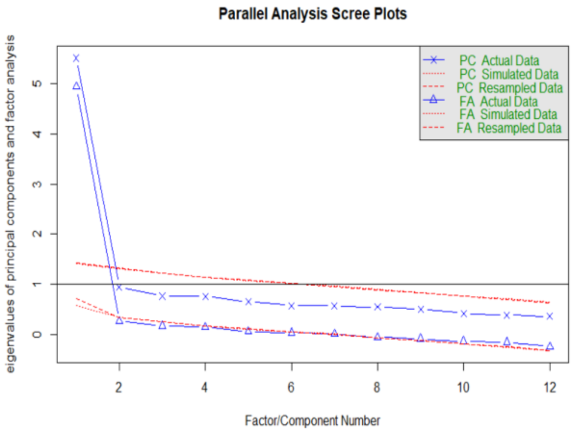
Appendix 4.2 Distribution of Responses per Item Score Category on the 5-Category Accuracy Scale

Items	No. of Responses (Percent)					
	0	1	2	3	4	Total
Form 1						
Q1	6(3.11%)	59(30.57%)	91(47.15%)	22(11.40%)	15(7.77%)	193(100%)
Q2	1(0.52%)	32(16.58%)	112(58.03%)	23(11.92%)	25(12.95%)	193(100%)
Q3	0(0%)	26(13.47%)	63(32.64%)	30(15.54%)	74(38.34%)	193(100%)
Q4	0(0%)	64(33.16%)	71(36.79%)	24(12.44%)	34(17.62%)	193(100%)
Q5	2(1.04%)	65(33.68%)	32(16.58%)	64(33.16%)	30(15.54%)	193(100%)
Q6	0(0%)	35(18.13%)	90(46.63%)	52(26.94%)	16(8.29%)	193(100%)
Q7	0(0%)	26(13.47%)	82(42.49%)	17(8.81%)	68(35.23%)	193(100%)
Q8	2(1.04%)	51(26.42%)	52(26.94%)	52(26.94%)	36(18.65%)	193(100%)
Q9	0(0%)	51(26.42%)	106(54.92%)	32(16.58%)	4(2.07%)	193(100%)
Q10	33(17.10%)	117(60.62%)	34(17.62%)	8(4.15%)	1(0.52%)	193(100%)
Q11	8(4.15%)	95(49.22%)	83(43.01%)	3(1.55%)	4(2.07%)	193(100%)
Q12	1(0.52%)	44(22.80%)	99(51.30%)	29(15.03%)	20(10.36%)	193(100%)
Form 2						
Q13	30(14.85%)	94(46.53%)	66(32.67%)	4(1.98%)	8(3.96%)	202(100%)
Q14	5(2.48%)	21(10.40%)	80(39.60%)	67(33.17%)	29(14.36%)	202(100%)
Q15	0(0%)	36(17.82%)	105(51.98%)	36(17.82%)	25(12.38%)	202(100%)
Q16	3(1.49%)	36(17.82%)	97(48.02%)	31(15.35%)	35(17.33%)	202(100%)
Q17	1(0.50%)	81(40.10%)	76(37.62%)	15(7.43%)	29(14.36%)	202(100%)
Q18	0(0%)	16(7.92%)	63(31.19%)	38(18.81%)	85(42.08%)	202(100%)
Q19	1(0.50%)	47(23.27%)	66(32.67%)	31(15.35%)	57(28.22%)	202(100%)
Q20	0(0%)	40(19.80%)	73(36.14%)	24(11.88%)	65(32.18%)	202(100%)
Q21	4(1.98%)	88(43.56%)	84(41.58%)	24(11.88%)	2(0.99%)	202(100%)
Q22	2(0.99%)	65(32.18%)	102(50.50%)	25(12.38%)	8(3.96%)	202(100%)
Q23	6(2.97%)	91(45.05%)	87(43.07%)	14(6.93%)	4(1.98%)	202(100%)
Q24	9(4.46%)	49(24.26%)	98(48.51%)	29(14.36%)	17(8.42%)	202(100%)
Form 3						
Q25	9(4.41%)	96(47.06%)	60(29.41%)	19(9.31%)	20(9.80%)	204(100%)
Q26	11(5.39%)	99(48.53%)	61(29.90%)	22(10.78%)	11(5.39%)	204(100%)
Q27	5(2.45%)	23(11.27%)	117(57.35%)	39(19.12%)	20(9.80%)	204(100%)
Q28	1(0.49%)	51(25.00%)	102(50.00%)	47(23.04%)	3(1.47%)	204(100%)
Q29	1(0.49%)	42(20.59%)	99(48.53%)	9(4.41%)	53(25.98%)	204(100%)
Q30	0(0%)	58(28.43%)	58(28.43%)	8(3.92%)	80(39.22%)	204(100%)
Q31	0(0%)	26(12.75%)	75(36.76%)	35(17.16%)	68(33.33%)	204(100%)
Q32	0(0%)	16(7.84%)	77(37.75%)	25(12.25%)	86(42.16%)	204(100%)

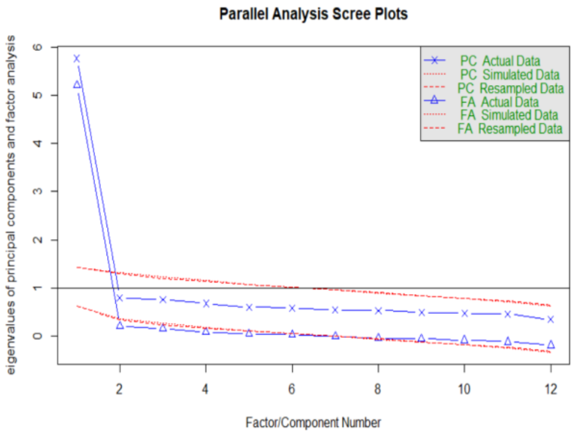
Q33	1 (0.49%)	40 (19.61%)	94 (46.08%)	40 (19.61%)	29 (14.22%)	204 (100%)
Q34	2 (0.98%)	56 (27.45%)	115 (56.37%)	23 (11.27%)	8 (3.92%)	204 (100%)
Q35	2 (0.98%)	28 (13.73%)	109 (53.43%)	27 (13.24%)	38 (18.63%)	204 (100%)
Q36	3 (1.47%)	79 (38.73%)	96 (47.06%)	21 (10.29%)	5 (2.45%)	204 (100%)
Form 4						
Q37	1 (0.56%)	61 (33.89%)	83 (46.11%)	20 (11.11%)	15 (8.33%)	180 (100%)
Q38	0 (0%)	36 (20%)	90 (50%)	25 (13.89%)	29 (16.11%)	180 (100%)
Q39	6 (3.33%)	69 (38.33%)	73 (40.56%)	19 (10.56%)	13 (7.22%)	180 (100%)
Q40	0 (0%)	48 (26.67%)	102 (56.67%)	27 (15.00%)	3 (1.67%)	180 (100%)
Q41	5 (2.78%)	34 (18.89%)	67 (37.22%)	18 (10%)	56 (31.11%)	180 (100%)
Q42	0 (0%)	44 (24.44%)	83 (46.11%)	21 (11.67%)	32 (17.78%)	180 (100%)
Q43	1 (0.56%)	45 (25.00%)	84 (46.67%)	38 (21.11%)	12 (6.67%)	180 (100%)
Q44	0 (0%)	15 (8.33%)	76 (42.22%)	25 (13.89%)	64 (35.56%)	180 (100%)
Q45	4 (2.22%)	90 (50%)	75 (41.67%)	9 (5.00%)	2 (1.11%)	180 (100%)
Q46	1 (0.56%)	30 (16.67%)	93 (51.67%)	23 (12.78%)	33 (18.33%)	180 (100%)
Q47	1 (0.56%)	40 (22.22%)	102 (56.67%)	29 (16.11%)	8 (4.44%)	180 (100%)
Q48	0 (0%)	11 (6.11%)	81 (45.00%)	60 (33.33%)	28 (15.56%)	180 (100%)

Appendix 4.3 Results of Parallel analyses: Scores Measured on Three EI Accuracy Scales

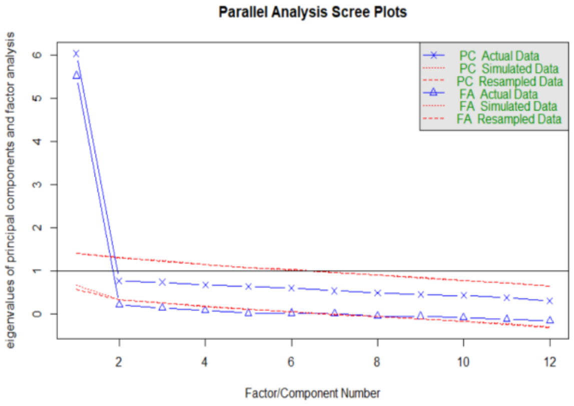
5-Category Scale (Form 1, Item 1 to 12)



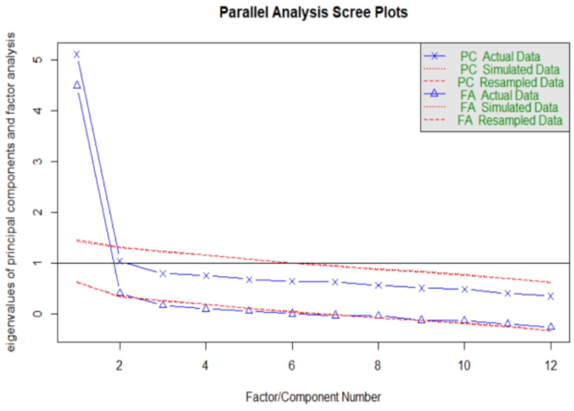
5-Category Scale (Form 2, Item 13 to 24)



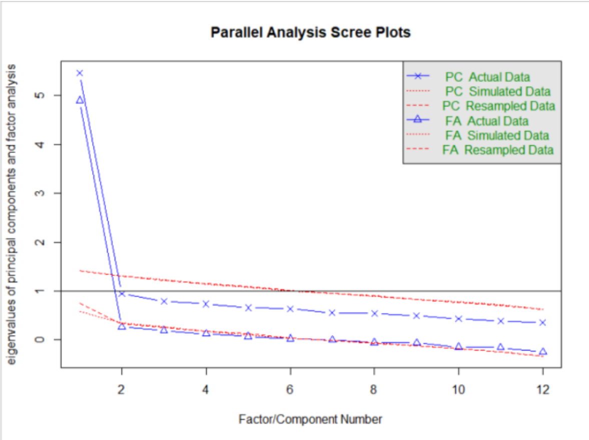
5-Category Scale (Form 3, Item 25 to 36)



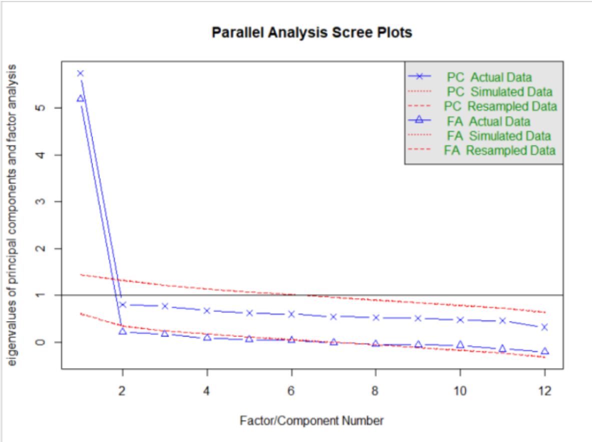
5-Category Scale (Form 4, Item 37 to 48)



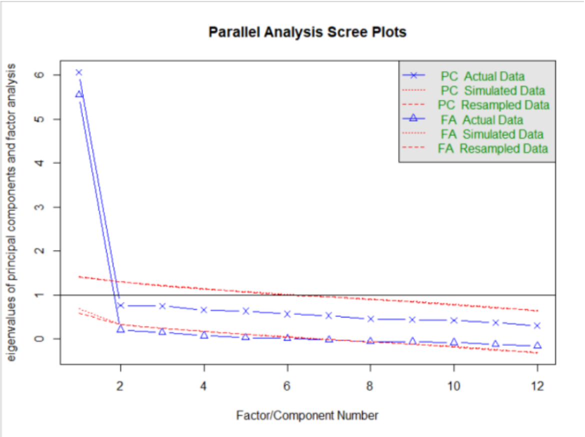
4-Category Scale (Form 1, Item 1 to 12)



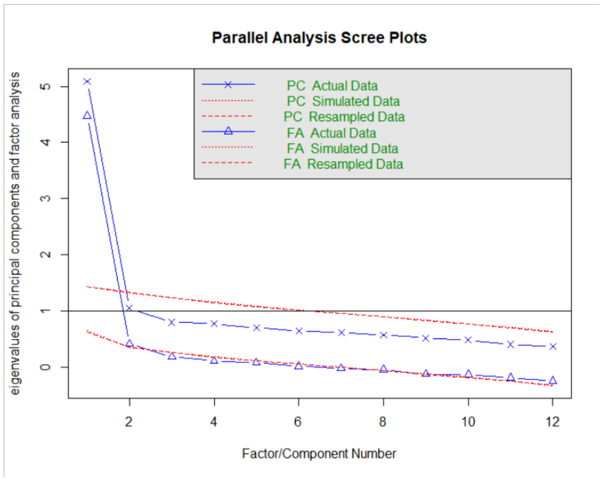
4-Category Scale (Form 2, Item 13 to 24)



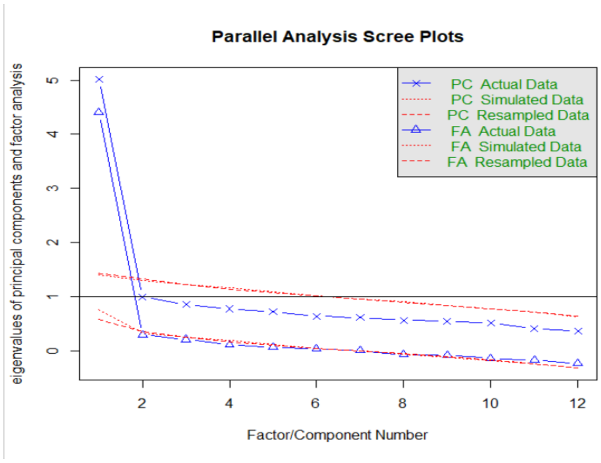
4-Category Scale (Form 3, Item 25 to 36)



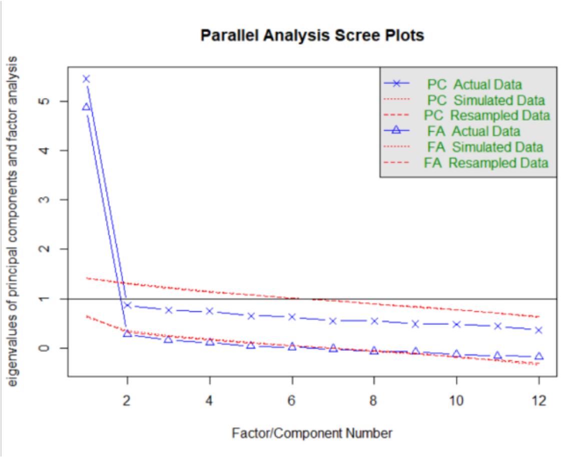
4-Category Scale (Form 4, Item 37 to 48)



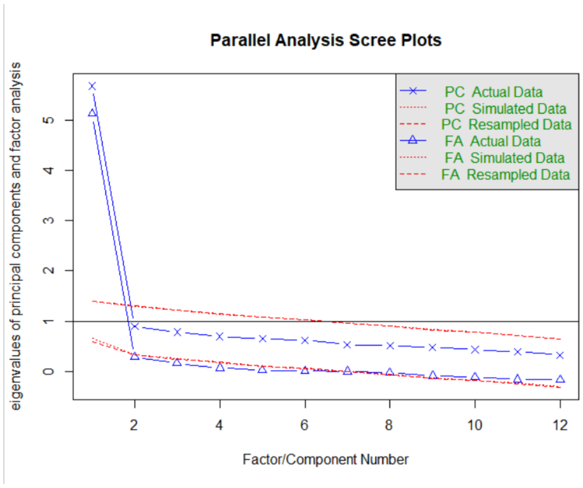
3-Category Scale (Form 1, Item 1 to 12)



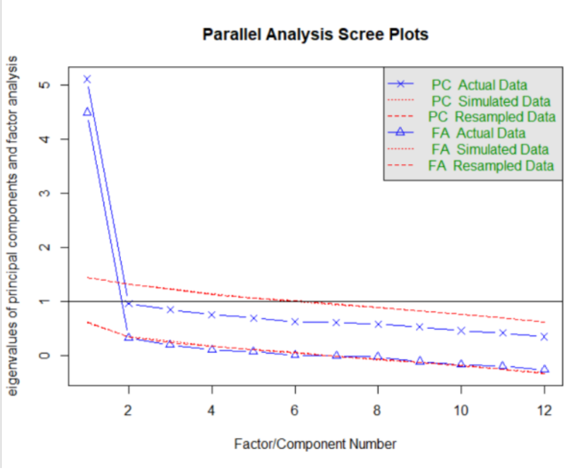
3-Category Scale (Form 2, Item 13 to 24)



3-Category Scale (Form 3, Item 25 to 36)



3-Category Scale (Form 4, Item 37 to 48)



Appendix 4.4. Results of Testing the Assumption of Local Independence: Pairwise LD-X^2 values

Item	**LD**											
5-Category Scale												
Form 1	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Q1	NA											
Q2	6.28	NA										
Q3	0.41	3.37	NA									
Q4	3.56	3.18	4.94	NA								
Q5	6.38	5.49	1.37	6.68	NA							
Q6	3.15	1.85	4.29	2.60	1.60	NA						
Q7	4.05	1.88	5.32	2.06	0.80	-0.01	NA					
Q8	1.31	6.18	2.52	4.34	6.81	1.90	6.48	NA				
Q9	3.89	1.70	3.71	2.75	1.88	0.61	0.66	5.51	NA			
Q10	2.59	3.37	0.63	2.16	0.94	4.26	2.01	5.19	3.19	NA		
Q11	2.34	1.37	5.04	6.34	4.27	3.85	1.04	2.32	1.05	3.11	NA	
Q12	5.13	7.17	5.71	2.89	5.47	1.32	1.64	8.39	0.57	0.54	6.69	NA
Form 2	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24
Q13	NA											
Q14	3.47	NA										
Q15	3.16	4.81	NA									
Q16	4.95	6.86	1.62	NA								
Q17	1.61	6.63	2.31	2.35	NA							
Q18	4.17	4.04	2.80	4.71	4.51	NA						
Q19	7.41	2.42	4.27	5.01	1.37	8.19	NA					
Q20	4.25	2.88	0.69	0.35	2.11	1.78	3.64	NA				
Q21	6.08	4.47	1.05	4.93	2.54	5.37	6.13	3.30	NA			
Q22	12.46	9.92	3.97	3.23	3.53	1.72	6.24	2.03	11.48	NA		
Q23	2.63	5.95	1.99	2.91	0.15	3.07	3.65	2.74	4.12	5.40	NA	
Q24	2.30	5.03	3.39	4.57	6.22	4.14	4.10	10.39	2.41	6.85	1.48	NA
Form 3	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36
Q25	NA											
Q26	3.70	NA										
Q27	11.81	1.86	NA									
Q28	8.79	3.97	6.80	NA								
Q29	8.55	2.15	3.16	3.43	NA							
Q30	4.96	4.11	0.85	4.06	8.45	NA						
Q31	7.22	6.48	1.98	1.25	9.46	2.05	NA					

Q32	8.71	4.06	2.18	5.27	3.42	5.44	5.77	NA				
Q33	9.02	4.54	6.23	20.67	14.74	2.54	0.73	3.09	NA			
Q34	2.76	2.70	1.93	8.99	4.45	13.25	2.64	4.13	7.38	NA		
Q35	6.05	3.18	7.94	3.08	3.06	3.44	2.44	3.93	6.05	3.76	NA	
Q36	4.52	2.81	4.75	3.70	4.94	2.69	1.53	5.09	9.36	4.09	3.10	NA

Form 4	Q37	Q38	Q39	Q40	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48
Q37	NA											
Q38	3.71	NA										
Q39	2.69	3.09	NA									
Q40	1.55	2.97	3.72	NA								
Q41	8.68	2.06	4.43	1.67	NA							
Q42	2.23	-0.57	1.89	-0.16	-0.03	NA						
Q43	4.71	3.24	34.80	3.38	3.23	2.84	NA					
Q44	0.66	1.97	4.65	-0.59	4.78	2.86	5.91	NA				
Q45	0.08	0.85	5.58	-0.34	1.38	1.60	1.89	0.66	NA			
Q46	0.80	1.46	5.65	1.54	1.65	3.86	2.49	0.19	0.85	NA		
Q47	2.30	2.17	2.46	0.86	0.70	2.32	1.16	1.73	0.23	0.90	NA	
Q48	2.05	2.91	1.10	1.52	4.85	1.86	2.54	5.40	1.15	1.44	3.46	NA

4-Category Scale

Form 1	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Q1	NA											
Q2	3.10	NA										
Q3	-0.32	2.85	NA									
Q4	2.75	2.96	4.92	NA								
Q5	3.38	2.46	1.20	2.93	NA							
Q6	2.83	1.50	4.23	2.70	1.08	NA						
Q7	3.33	1.52	5.33	2.05	-0.06	-0.05	NA					
Q8	-0.39	1.89	2.25	3.85	-0.11	1.01	6.29	NA				
Q9	2.83	1.58	3.54	2.72	0.07	0.70	0.64	5.38	NA			
Q10	0.24	2.51	0.22	1.19	0.38	1.28	0.47	1.33	2.51	NA		
Q11	1.83	0.27	2.28	5.64	0.72	2.15	0.24	-0.04	0.79	0.73	NA	
Q12	4.39	3.52	5.17	2.72	2.41	0.84	1.26	2.77	0.44	-0.11	4.91	NA

Form 2	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24
Q13	NA											
Q14	2.50	NA										
Q15	0.57	1.63	NA									
Q16	0.70	3.18	1.35	NA								
Q17	0.36	3.25	2.04	1.84	NA							

Q18	2.74	1.72	2.81	3.71	4.07	NA						
Q19	2.48	0.59	3.89	3.71	1.13	7.12	NA					
Q20	6.22	2.01	0.68	0.21	1.23	1.77	2.72	NA				
Q21	3.15	0.56	0.86	2.45	1.87	3.30	3.33	2.90	NA			
Q22	12.01	6.45	3.95	2.00	3.12	0.16	1.21	1.69	5.11	NA		
Q23	1.54	4.94	1.26	0.54	-0.38	0.94	1.36	2.00	2.20	0.32	NA	
Q24	0.24	2.04	1.45	1.63	4.73	3.30	3.11	1.69	1.78	3.08	1.06	NA
Form 3	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36
Q25	NA											
Q26	1.35	NA										
Q27	5.62	0.59	NA									
Q28	3.92	2.53	3.70	NA								
Q29	3.10	0.47	2.48	2.60	NA							
Q30	3.00	2.07	0.70	3.84	5.65	NA						
Q31	6.06	4.02	1.80	0.68	6.73	2.41	NA					
Q32	7.85	2.13	1.82	4.40	3.22	5.29	5.80	NA				
Q33	5.16	3.43	3.08	1.79	4.78	2.32	0.50	2.61	NA			
Q34	1.89	2.04	1.60	3.14	3.92	4.07	2.27	1.47	5.40	NA		
Q35	4.20	0.80	2.05	2.37	2.08	2.98	0.37	3.69	4.37	2.85	NA	
Q36	3.90	1.05	3.36	3.40	4.51	2.36	0.90	4.52	2.14	1.50	1.68	NA
Form 4	Q37	Q38	Q39	Q40	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48
Q37	NA											
Q38	3.24	NA										
Q39	0.70	4.52	NA									
Q40	1.05	2.94	2.77	NA								
Q41	3.71	1.78	3.01	1.48	NA							
Q42	1.85	-0.62	1.77	-0.17	-0.30	NA						
Q43	4.19	2.42	4.14	2.62	1.98	2.59	NA					
Q44	0.30	2.02	1.77	-0.54	3.61	2.87	0.89	NA				
Q45	-0.31	0.56	5.06	-0.57	0.29	0.71	1.51	-0.14	NA			
Q46	0.23	0.82	2.87	1.39	0.45	3.60	0.63	-0.30	0.75	NA		
Q47	1.66	1.60	2.06	0.67	-0.34	2.11	0.44	1.16	-0.39	0.20	NA	
Q48	0.95	2.94	0.70	1.51	3.64	1.85	2.21	5.28	1.04	1.11	2.58	NA
3-Category Scale												
Form 1	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Q1	NA											
Q2	-0.47	NA										
Q3	-1.04	2.12	NA									

Q4	0.30	1.72	2.30	NA								
Q5	0.92	1.05	0.27	1.62	NA							
Q6	-0.14	0.92	2.81	-0.16	-0.08	NA						
Q7	1.17	0.39	2.05	-0.09	-0.82	-0.72	NA					
Q8	-1.14	0.10	1.09	-0.02	-1.17	-0.79	2.60	NA				
Q9	-0.57	-0.07	2.18	0.54	-1.09	-0.53	-0.11	-0.68	NA			
Q10	-0.91	-0.53	-0.22	0.02	-0.44	-0.72	0.16	0.43	-1.13	NA		
Q11	-0.41	-0.46	-0.08	-0.52	-0.95	0.65	-0.62	-0.82	-0.75	-0.74	NA	
Q12	2.24	1.15	2.25	-0.07	0.53	-1.20	-0.20	-0.30	-0.96	-0.84	0.68	NA
Form 2	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24
Q13	NA											
Q14	-1.30	NA										
Q15	-0.41	-0.11	NA									
Q16	-0.86	-0.03	0.49	NA								
Q17	-0.74	-0.99	0.37	0.78	NA							
Q18	-1.17	-0.63	1.05	2.11	-0.46	NA						
Q19	-0.13	-0.67	1.51	1.27	0.14	0.61	NA					
Q20	2.50	-0.52	-1.24	-0.47	0.01	-1.16	0.28	NA				
Q21	2.04	-0.46	-0.80	1.61	0.53	1.06	-0.18	1.67	NA			
Q22	4.55	4.25	1.21	1.22	1.19	-0.69	-0.04	0.38	0.48	NA		
Q23	-0.50	1.65	0.43	-0.87	-0.96	-0.34	0.46	-0.35	-0.49	-0.37	NA	
Q24	-1.25	-0.90	-0.58	0.56	0.44	-0.70	0.59	-0.18	0.03	1.78	-0.59	NA
Form 3	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36
Q25	NA											
Q26	0.22	NA										
Q27	2.27	-0.07	NA									
Q28	3.47	0.98	3.18	NA								
Q29	2.66	0.04	1.94	2.82	NA							
Q30	1.63	1.49	-0.86	1.37	2.83	NA						
Q31	2.05	0.84	-0.05	-0.68	3.68	-0.09	NA					
Q32	7.51	0.92	0.70	3.71	2.27	4.89	3.65	NA				
Q33	2.59	-0.38	0.11	1.56	4.60	1.68	-0.68	0.74	NA			
Q34	0.31	-0.18	-0.24	1.32	2.98	0.94	0.92	1.26	2.93	NA		
Q35	0.72	-0.07	-0.24	1.79	1.01	0.98	-0.98	1.89	1.92	2.24	NA	
Q36	2.42	0.16	-0.19	1.12	3.42	1.79	-0.21	4.11	1.10	0.91	0.67	NA
Form 4	Q37	Q38	Q39	Q40	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48
Q37	NA											
Q38	0.15	NA										

Q39	-0.02	2.04	NA									
Q40	-1.17	0.00	0.99	NA								
Q41	1.81	0.24	1.06	0.06	NA							
Q42	-0.80	-0.83	1.28	-0.80	-0.76	NA						
Q43	-0.98	-0.36	1.81	0.87	0.49	-0.30	NA					
Q44	0.25	-0.04	-0.21	-0.62	-0.31	-0.59	-0.08	NA				
Q45	-0.46	-0.96	1.41	-0.89	-0.75	-0.30	0.47	-0.56	NA			
Q46	-1.04	-1.24	-0.57	0.10	-0.66	-0.46	-0.92	-0.91	0.46	NA		
Q47	-0.23	0.35	-0.22	-0.33	-1.30	-0.30	0.43	-0.19	-1.08	-0.43	NA	
Q48	-0.88	-0.27	-0.66	-0.53	2.82	-0.28	-0.12	2.29	-0.02	-0.44	1.59	NA

Appendix 4.5 Overall Model Fit and Factor Loadings of the GRM Model Fitted to EI Scores on the Three Accuracy Scales

Model	**Number of Parameters**	**M^2**	***df* (M^2)**	***p* (M^2)**	**Log -likelihood**	**RMSEA (M^2)**	**SRMR**	**TLI**	**CFI**	**AIC**	**BIC**
Form 1 (Q1 to Q12)											
5-cat	55	25.51	23	0.32	-2467.43	0.02	0.05	0.96	0.98	5044.86	5224.31
4-cat.	48	27.71	30	0.59	-2327.16	<0.01	0.05	1.01	1.00	4750.32	4906.93
3-cat.	36	43.48	42	0.41	-1946.02	0.01	0.05	1.00	1.00	3964.05	4081.50
Form 2 (Q13 to Q24)											
5-cat.	57	11.51	21	0.95	-2592.80	<0.01	0.04	1.10	1.00	5299.60	5488.18
4-cat.	48	14.98	30	0.99	-2427.00	<0.01	0.04	1.03	1.00	4949.99	5108.79
3-cat.	36	30.72	42	0.90	-2014.92	<0.01	0.04	1.01	1.00	4101.83	4220.93
Form 3 (Q25 to Q36)											
5-cat.	57	12.21	21	0.93	-2498.23	<0.01	0.04	1.06	1.00	5110.46	5299.60
4-cat.	48	20.58	30	0.90	-2378.40	<0.01	0.05	1.01	1.00	4852.80	5012.07
3-cat.	36	41.78	42	0.48	-2018.99	<0.01	0.05	1.00	1.00	4109.98	4229.44
Form 4 (Q37 to Q48)											
5-cat.	55	40.48	23	0.01	-2297.77	0.07	0.06	0.88	0.92	4705.54	4881.16
4-cat.	48	43.81	30	0.05*	-2230.88	0.05	0.06	0.96	0.97	4557.76	4711.03
3-cat.	36	53.36	42	0.11	-1878.74	0.04	0.05	0.99	0.99	3829.49	3944.43

Note. 5-cat. = a graded response model fitted to EI item scores on the original 5-category accuracy scale; 4-cat. = a GRM model fitted to EI item scores on the 4-category scale that collapsed the item score 0 into the item score 1 category of the original scale; 3-cat. = a GRM model fitted to EI item scores on the 3-cateogry scale that collapsed the item score 4 into item score 3 category of the 4-category scale; M^2 = model fit statistic; df = degree of freedom; *p* = *p*-value associated with model-fit statistic; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; TLI = Tucker–Lewis index; CFI = Comparative Fit Index; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; *p = 0.0496

Appendix 4.6 Factor Loadings of Exploratory Factor Analysis for the 48 EI Items Measured on the Three EI Accuracy Scales

Item	5-category scale		4-category scale		3-category scale	
	loading	h2	loading	h2	loading	h2
Form 1						
Q1	0.72	0.51	0.71	0.50	0.72	0.51
Q2	0.76	0.58	0.77	0.59	0.78	0.61
Q3	0.62	0.39	0.61	0.38	0.59	0.35
Q4	0.68	0.46	0.68	0.47	0.71	0.51
Q5	0.67	0.45	0.69	0.48	0.69	0.47
Q6	0.57	0.33	0.59	0.34	0.59	0.35
Q7	0.58	0.34	0.58	0.34	0.57	0.32
Q8	0.71	0.50	0.72	0.52	0.70	0.49
Q9	0.79	0.62	0.79	0.62	0.81	0.66
Q10	0.72	0.52	0.68	0.47	0.70	0.49
Q11	0.48	0.23	0.51	0.26	0.50	0.25
Q12	0.73	0.54	0.74	0.54	0.70	0.49
SS loadings	5.47		5.50		5.50	
Var. Explained	45.6%		45.8%		45.9%	
Form 2						
Q13	0.79	0.62	0.84	0.71	0.84	0.70
Q14	0.65	0.42	0.65	0.42	0.74	0.54
Q15	0.69	0.47	0.69	0.48	0.65	0.42
Q16	0.70	0.49	0.71	0.50	0.69	0.48
Q17	0.74	0.54	0.73	0.54	0.74	0.55
Q18	0.67	0.45	0.66	0.43	0.78	0.60
Q19	0.68	0.46	0.67	0.45	0.64	0.40
Q20	0.78	0.60	0.78	0.61	0.77	0.60
Q21	0.66	0.43	0.65	0.42	0.62	0.39
Q22	0.64	0.41	0.62	0.38	0.61	0.37
Q23	0.63	0.40	0.61	0.38	0.62	0.39
Q24	0.79	0.63	0.81	0.66	0.82	0.67
SS loadings	5.91		5.97		6.11	
Var. Explained	49.2%		49.7%		50.9%	
Form 3						
Q25	0.78	0.60	0.80	0.64	0.81	0.65
Q26	0.74	0.55	0.76	0.58	0.76	0.58
Q27	0.67	0.45	0.67	0.45	0.66	0.44

Q28	0.64	0.42	0.65	0.42	0.63	0.40
Q29	0.77	0.59	0.78	0.61	0.79	0.62
Q30	0.64	0.41	0.65	0.42	0.64	0.41
Q31	0.65	0.43	0.66	0.43	0.62	0.38
Q32	0.67	0.44	0.67	0.45	0.71	0.51
Q33	0.77	0.59	0.75	0.57	0.79	0.63
Q34	0.84	0.71	0.83	0.69	0.83	0.69
Q35	0.83	0.69	0.83	0.68	0.86	0.74
Q36	0.64	0.41	0.63	0.39	0.61	0.37
SS loadings	6.28		6.33		6.41	
Var. Explained	52.3%		52.7%		53.4%	
Form 4						
Q37	0.52	0.27	0.53	0.28	0.51	0.26
Q38	0.71	0.50	0.70	0.50	0.70	0.49
Q39	0.62	0.39	0.63	0.40	0.66	0.44
Q40	0.65	0.42	0.65	0.42	0.66	0.44
Q41	0.67	0.44	0.66	0.43	0.64	0.41
Q42	0.70	0.49	0.70	0.49	0.73	0.53
Q43	0.68	0.46	0.69	0.48	0.73	0.53
Q44	0.63	0.39	0.64	0.41	0.68	0.46
Q45	0.67	0.45	0.65	0.43	0.60	0.36
Q46	0.72	0.51	0.72	0.52	0.70	0.49
Q47	0.71	0.50	0.70	0.49	0.72	0.52
Q48	0.56	0.31	0.56	0.31	0.60	0.36
SS loadings	5.13		5.16		5.30	
Var. Explained	42.7%		43.0%		44.1%	

Note: SS loadings = sum of standardized factor loadings; Var. explained=the proportion of the variance explained

Appendix 4.7 The Item Discrimination (a) and Item Difficulty (b) of the 48 EI Items Measured on the 5-Category, 4-Category, and 3-Category EI Accuracy Scales

Item		5-category scale						4-category scale					3-category scale			
		a	**b_1**	**b_2**	**b_3**	**b_4**	**$b_{overall}$**	**a**	**b_1**	**b_2**	**b_3**	**$b_{overall}$**	**a**	**b_1**	**b_2**	**$b_{overall}$**
From 1																
Q1	Coef.	1.74	-2.67	-0.64	1.21	2.00	-0.02	1.70	-0.64	1.22	2.03	0.87	1.74	-0.63	1.23	0.30
	SE	0.25	0.36	0.14	0.18	0.26		0.25	0.14	0.18	0.27		0.26	0.14	0.18	
Q2	Coef.	2.00	-3.43	-1.29	0.91	1.55	-0.57	2.03	-1.28	0.90	1.54	0.39	2.11	-1.25	0.87	-0.19
	SE	0.29	0.61	0.17	0.14	0.19		0.29	0.17	0.14	0.19		0.32	0.16	0.14	
Q3	Coef.	1.35	-1.80	-0.22	0.45	N/A	-0.52	1.33	-1.81	-0.22	0.46	-0.53	1.25	-1.87	-0.20	-1.04
	SE	0.21	0.27	0.14	0.16	N/A		0.21	0.28	0.15	0.16		0.22	0.31	0.15	
Q4	Coef.	1.58	-0.69	0.76	1.41	N/A	0.49	1.59	-0.69	0.76	1.41	0.49	1.73	-0.66	0.69	0.01
	SE	0.23	0.15	0.15	0.20	N/A		0.23	0.15	0.15	0.20		0.26	0.14	0.15	
Q5	Coef.	1.53	-3.60	-0.63	0.02	1.56	-0.66	1.62	-0.60	0.01	1.51	0.31	1.61	-0.60	0.02	-0.29
	SE	0.22	0.61	0.15	0.13	0.22		0.23	0.14	0.13	0.21		0.26	0.14	0.13	
Q6	Coef.	1.19	-1.62	0.64	2.47	N/A	0.49	1.23	-1.58	0.63	2.41	0.49	1.24	-1.58	0.62	-0.48
	SE	0.19	0.26	0.18	0.38	N/A		0.20	0.25	0.17	0.37		0.21	0.26	0.17	
Q7	Coef.	1.22	-1.91	0.24	0.64	N/A	-0.34	1.21	-1.91	0.24	0.64	-0.34	1.18	-1.95	0.23	-0.86
	SE	0.20	0.30	0.16	0.18	N/A		0.20	0.31	0.16	0.18		0.21	0.33	0.16	
Q8	Coef.	1.70	-3.36	-0.88	0.11	1.27	-0.71	1.77	-0.87	0.11	1.25	0.17	1.67	-0.88	0.13	-0.38
	SE	0.24	0.55	0.15	0.13	0.18		0.25	0.15	0.12	0.18		0.27	0.16	0.13	
Q9	Coef.	2.19	-0.83	1.20	2.67	N/A	1.01	2.18	-0.83	1.20	2.67	1.01	2.35	-0.79	1.14	0.18
	SE	0.32	0.13	0.15	0.34	N/A		0.32	0.13	0.16	0.34		0.36	0.13	0.15	
Q10	Coef.	1.77	-1.34	1.08	2.38	3.78	1.48	1.60	1.12	2.52	4.04	2.56	1.68	1.09	2.48	1.78
	SE	0.26	0.18	0.16	0.31	0.70		0.30	0.18	0.39	0.84		0.32	0.18	0.36	
Q11	Coef.	0.94	-3.74	0.14	3.91	4.56	1.22	1.02	0.15	3.67	4.27	2.70	0.99	0.15	3.77	1.96
	SE	0.19	0.74	0.18	0.77	0.94		0.20	0.17	0.71	0.87		0.20	0.18	0.74	

Q12	Coef.	1.84	-3.64	-1.02	0.89	1.77	-0.50	1.85	-1.01	0.89	1.76	0.55	1.68	-1.06	0.93	-0.06
	SE	0.25	0.66	0.16	0.15	0.22	N/A	0.26	0.16	0.15	0.22		0.25	0.17	0.16	
From 2																
Q13	Coef.	2.15	-1.35	0.34	2.05	2.33	0.84	2.64	0.33	1.90	2.14	1.46	2.61	0.33	1.99	1.16
	SE	0.28	0.16	0.12	0.23	0.27		0.41	0.11	0.21	0.24		0.42	0.11	0.22	
Q14	Coef.	1.44	-3.19	-1.77	0.12	1.71	-0.78	1.44	-1.76	0.12	1.71	0.02	1.85	-1.54	0.06	-0.74
	SE	0.20	0.47	0.24	0.14	0.24		0.20	0.24	0.14	0.24		0.28	0.20	0.12	
Q15	Coef.	1.60	-1.37	0.71	1.72	N/A	0.36	1.64	-1.35	0.71	1.69	0.35	1.46	-1.41	0.82	-0.30
	SE	0.22	0.19	0.15	0.23	N/A		0.23	0.19	0.15	0.22		0.22	0.21	0.17	
Q16	Coef.	1.68	-3.13	-1.25	0.63	1.38	-0.59	1.69	-1.25	0.63	1.38	0.25	1.64	-1.27	0.65	-0.31
	SE	0.23	0.48	0.18	0.14	0.19		0.23	0.18	0.14	0.19		0.24	0.19	0.15	
Q17	Coef.	1.85	-3.67	-0.35	1.05	1.46	-0.38	1.83	-0.35	1.05	1.47	0.72	1.87	-0.35	1.05	0.35
	SE	0.25	0.66	0.12	0.16	0.19		0.26	0.12	0.16	0.19		0.27	0.12	0.16	
Q18	Coef.	1.52	-2.16	-0.38	0.34	N/A	-0.73	1.49	-2.20	-0.39	0.35	-0.75	2.09	-1.80	-0.39	-1.09
	SE	0.22	0.29	0.14	0.14	N/A		0.22	0.30	0.14	0.14		0.34	0.22	0.12	
Q19	Coef.	1.56	-4.08	-1.09	0.16	0.81	-1.05	1.54	-1.11	0.17	0.82	-0.04	1.40	-1.16	0.21	-0.48
	SE	0.22	0.78	0.17	0.13	0.16		0.22	0.18	0.13	0.16		0.22	0.19	0.14	
Q20	Coef.	2.10	-1.12	0.14	0.57	N/A	-0.13	2.12	-1.11	0.15	0.57	-0.13	2.07	-1.10	0.16	-0.47
	SE	0.28	0.15	0.11	0.13	N/A		0.29	0.15	0.11	0.13		0.30	0.15	0.12	
Q21	Coef.	1.47	-3.27	-0.22	1.69	3.86	0.52	1.46	-0.22	1.70	3.89	1.79	1.36	-0.21	1.79	0.79
	SE	0.22	0.50	0.13	0.24	0.65		0.22	0.14	0.25	0.66		0.22	0.14	0.27	
Q22	Coef.	1.42	-3.85	-0.71	1.50	2.89	-0.04	1.34	-0.75	1.55	2.99	1.26	1.30	-0.76	1.65	0.44
	SE	0.21	0.67	0.16	0.22	0.41		0.21	0.16	0.24	0.44		0.21	0.17	0.26	
Q23	Coef.	1.39	-3.06	-0.11	2.14	3.46	0.61	1.32	-0.11	2.21	3.57	1.89	1.35	-0.09	2.20	1.06
	SE	0.21	0.47	0.14	0.31	0.54		0.22	0.14	0.33	0.59		0.22	0.14	0.33	
Q24	Coef.	2.22	-2.16	-0.75	0.95	1.80	-0.04	2.35	-0.71	0.93	1.76	0.66	2.44	-0.70	0.96	0.13
	SE	0.28	0.25	0.12	0.14	0.20		0.32	0.12	0.13	0.19		0.35	0.12	0.14	

From 3																
Q25	Coef.	2.10	-2.12	-0.05	1.17	1.80	0.20	2.25	-0.04	1.14	1.75	0.95	2.33	-0.06	1.09	0.51
	SE	0.27	0.26	0.11	0.15	0.20		0.30	0.11	0.15	0.19		0.33	0.11	0.15	
Q26	Coef.	1.87	-2.15	0.05	1.36	2.33	0.40	2.01	0.05	1.31	2.25	1.20	1.99	0.04	1.30	0.67
	SE	0.24	0.27	0.12	0.17	0.27		0.27	0.11	0.17	0.25		0.28	0.12	0.18	
Q27	Coef.	1.53	-2.98	-1.63	0.82	2.03	-0.44	1.53	-1.63	0.82	2.03	0.40	1.51	-1.55	0.82	-0.37
	SE	0.21	0.44	0.22	0.16	0.26		0.21	0.22	0.16	0.26		0.23	0.22	0.16	
Q28	Coef.	1.43	-4.29	-1.06	1.05	3.66	-0.16	1.45	-1.05	1.04	3.62	1.20	1.40	-1.10	1.05	-0.03
	SE	0.21	0.87	0.17	0.18	0.57		0.21	0.17	0.18	0.56		0.21	0.18	0.19	
Q29	Coef.	2.06	-3.35	-1.09	0.63	0.84	-0.74	2.12	-1.07	0.63	0.83	0.13	2.19	-1.05	0.60	-0.22
	SE	0.28	0.60	0.14	0.13	0.14		0.29	0.14	0.13	0.14		0.31	0.14	0.13	
Q30	Coef.	1.43	-0.92	0.20	0.37	N/A	-0.12	1.44	-0.91	0.21	0.35	-0.12	1.41	-0.89	0.20	-0.35
	SE	0.22	0.17	0.14	0.15	N/A		0.22	0.17	0.14	0.15		0.23	0.17	0.14	
Q31	Coef.	1.47	-1.73	-0.08	0.61	N/A	-0.40	1.49	-1.72	-0.08	0.61	-0.40	1.34	-1.83	-0.06	-0.94
	SE	0.21	0.24	0.13	0.16	N/A		0.21	0.24	0.13	0.15		0.23	0.28	0.14	
Q32	Coef.	1.52	-2.11	-0.20	0.28	N/A	-0.68	1.53	-2.10	-0.20	0.28	-0.67	1.73	-1.95	-0.23	-1.09
	SE	0.22	0.30	0.13	0.14	N/A		0.23	0.29	0.13	0.14		0.28	0.27	0.12	
Q33	Coef.	2.02	-3.39	-1.11	0.53	1.51	-0.62	1.95	-1.14	0.54	1.53	0.31	2.21	-1.08	0.48	-0.30
	SE	0.26	0.60	0.15	0.13	0.18		0.25	0.15	0.13	0.18		0.32	0.14	0.12	
Q34	Coef.	2.66	-2.64	-0.74	1.29	2.24	0.04	2.56	-0.74	1.29	2.26	0.94	2.56	-0.74	1.29	0.28
	SE	0.38	0.37	0.11	0.15	0.23		0.38	0.12	0.15	0.24		0.39	0.12	0.15	
Q35	Coef.	2.51	-2.72	-1.26	0.55	1.13	-0.57	2.50	-1.26	0.56	1.13	0.14	2.83	-1.21	0.51	-0.35
	SE	0.33	0.38	0.14	0.12	0.14		0.33	0.15	0.12	0.14		0.43	0.14	0.12	
Q36	Coef.	1.41	-3.53	-0.46	1.80	3.29	0.28	1.37	-0.46	1.83	3.35	1.57	1.29	-0.46	1.90	0.72
	SE	0.21	0.59	0.14	0.25	0.49		0.21	0.14	0.26	0.51		0.21	0.15	0.29	

From 4																
Q37	Coef.	1.04	-5.44	-0.78	1.62	2.70	-0.48	1.05	-0.77	1.61	2.69	1.18	1.02	-0.75	1.70	0.47
	SE	0.19	1.31	0.21	0.31	0.48		0.20	0.21	0.31	0.48		0.20	0.22	0.33	
Q38	Coef.	1.70	-1.21	0.74	1.42	N/A	0.32	1.69	-1.21	0.75	1.42	0.32	1.67	-1.16	0.74	-0.21
	SE	0.26	0.19	0.15	0.21	N/A		0.26	0.19	0.15	0.21		0.27	0.19	0.16	
Q39	Coef.	1.35	-3.10	-0.32	1.55	2.43	0.14	1.39	-0.29	1.49	2.38	1.19	1.51	-0.25	1.42	0.59
	SE	0.21	0.48	0.15	0.24	0.36		0.23	0.15	0.23	0.36		0.25	0.14	0.22	
Q40	Coef.	1.45	-0.98	1.50	3.50	N/A	1.34	1.46	-0.97	1.49	3.48	1.33	1.51	-0.93	1.46	0.27
	SE	0.24	0.18	0.23	0.59	N/A		0.24	0.18	0.23	0.58		0.25	0.18	0.23	
Q41	Coef.	1.52	-2.98	-1.18	0.32	0.73	-0.78	1.49	-1.20	0.32	0.74	-0.05	1.42	-1.17	0.33	-0.42
	SE	0.24	0.46	0.20	0.14	0.17		0.24	0.20	0.14	0.17		0.24	0.21	0.15	
Q42	Coef.	1.67	-1.01	0.79	1.36	N/A	0.38	1.65	-1.01	0.79	1.37	0.38	1.80	-0.94	0.74	-0.10
	SE	0.25	0.17	0.16	0.20	N/A		0.25	0.17	0.16	0.20		0.28	0.16	0.15	
Q43	Coef.	1.57	-4.04	-0.99	0.91	2.26	-0.46	1.63	-0.96	0.90	2.21	0.72	1.82	-0.91	0.82	-0.05
	SE	0.24	0.79	0.17	0.17	0.32		0.25	0.17	0.16	0.31		0.29	0.16	0.16	
Q44	Coef.	1.37	-2.24	0.05	0.61	N/A	-0.53	1.41	-2.21	0.05	0.60	-0.52	1.58	-2.00	0.03	-0.98
	SE	0.23	0.34	0.15	0.17	N/A		0.23	0.33	0.14	0.16		0.27	0.29	0.14	
Q45	Coef.	1.53	-3.14	0.09	2.37	3.62	0.73	1.47	0.08	2.43	3.71	2.07	1.29	0.15	2.64	1.39
	SE	0.26	0.50	0.14	0.35	0.65		0.26	0.14	0.37	0.69		0.24	0.15	0.44	
Q46	Coef.	1.74	-3.73	-1.34	0.68	1.27	-0.78	1.78	-1.32	0.67	1.26	0.20	1.68	-1.35	0.70	-0.32
	SE	0.26	0.71	0.19	0.15	0.19		0.27	0.19	0.14	0.19		0.27	0.20	0.16	
Q47	Coef.	1.70	-3.82	-1.08	1.17	2.51	-0.30	1.68	-1.09	1.17	2.52	0.87	1.76	-0.99	1.17	0.09
	SE	0.26	0.73	0.18	0.18	0.34		0.26	0.18	0.18	0.35		0.28	0.17	0.18	
Q48	Coef.	1.14	-2.83	0.08	1.86	N/A	-0.30	1.15	-2.81	0.08	1.84	-0.30	1.27	-2.61	0.05	-1.28
	SE	0.20	0.48	0.16	0.32	N/A		0.20	0.48	0.16	0.31		0.24	0.45	0.15	

Note: Coef.=coefficients a and b; SE=standard error

Appendix 4.8 Item Fit Statistics of the 48 EI Items Measured on the Three EI Accuracy Scales

Item	5-Category Scale				4-Category Scale				3-Category Scale			
	$S\text{-}X^2$	*df*	RMSEA	*p*	$S\text{-}X^2$	*df*	RMSEA	*p*	$S\text{-}X^2$	*df*	RMSEA	*p*
Form 1												
Q1	29.95	26	0.03	0.27	30.73	27	0.03	0.28	18.68	18	0.01	0.41
Q2	45.00	26	0.06	0.012	37.68	27	0.05	0.08	8.38	13	<0.01	0.82
Q3	48.43	37	0.04	0.10	37.87	36	0.02	0.38	20.07	20	<0.01	0.45
Q4	37.99	35	0.02	0.34	39.36	34	0.03	0.24	19.75	19	0.01	0.41
Q5	31.18	33	<0.01	0.56	32.20	33	<0.01	0.51	21.63	21	0.01	0.42
Q6	25.10	33	<0.01	0.84	19.04	32	<0.01	0.97	16.40	21	<0.01	0.75
Q7	30.40	29	0.02	0.39	37.97	30	0.04	0.15	28.32	21	0.04	0.13
Q8	32.89	31	0.02	0.38	28.60	32	<0.01	0.64	19.18	19	0.01	0.45
Q9	15.16	19	<0.01	0.71	19.75	18	0.02	0.35	5.73	13	<0.01	0.96
Q10	16.68	22	<0.01	0.78	8.47	14	<0.01	0.86	8.29	10	<0.01	0.60
Q11	26.03	20	0.04	0.17	20.91	19	0.02	0.34	22.87	14	0.06	0.06
Q12	45.82	28	0.06	0.018	32.09	28	0.03	0.27	19.47	17	0.03	0.30
Form 2												
Q13	15.78	19	<0.01	0.67	11.92	13	<0.01	0.54	10.49	12	<0.01	0.57
Q14	31.32	32	<0.01	0.50	37.06	32	0.03	0.25	22.62	20	0.03	0.31
Q15	52.18	33	0.05	0.018	48.36	31	0.05	0.024	26.29	22	0.03	0.24
Q16	32.93	31	0.02	0.37	34.98	30	0.03	0.24	14.68	21	<0.01	0.84
Q17	17.15	28	<0.01	0.95	17.57	26	<0.01	0.89	21.44	19	0.03	0.31
Q18	40.63	29	0.05	0.07	39.89	27	0.05	0.05	17.31	14	0.03	0.24
Q19	60.37	38	0.05	0.012	37.23	36	0.01	0.41	19.95	24	<0.01	0.70
Q20	26.36	29	<0.01	0.61	33.16	29	0.03	0.27	17.44	20	<0.01	0.62
Q21	32.62	24	0.04	0.11	30.01	25	0.03	0.22	31.22	22	0.05	0.09
Q22	22.58	26	<0.01	0.66	24.06	24	<0.01	0.46	18.14	22	<0.01	0.70
Q23	32.14	18	0.06	0.021	29.66	19	0.05	0.06	20.93	19	0.02	0.34
Q24	26.33	27	<0.01	0.50	21.17	25	<0.01	0.68	15.60	14	0.02	0.34
Form3												
Q25	30.24	24	0.04	0.18	29.34	22	0.04	0.14	22.34	14	0.05	0.07
Q26	26.24	23	0.03	0.29	17.73	18	<0.01	0.47	11.72	16	<0.01	0.76
Q27	15.94	28	<0.01	0.97	18.55	30	<0.01	0.95	25.49	19	0.04	0.15
Q28	18.63	25	<0.01	0.82	21.61	26	<0.01	0.71	15.35	20	<0.01	0.76
Q29	20.28	20	0.01	0.44	22.28	17	0.04	0.17	33.87	17	0.07	0.009
Q30	31.27	27	0.03	0.26	35.65	25	0.05	0.08	21.20	20	0.02	0.39
Q31	41.52	34	0.03	0.18	40.53	36	0.03	0.28	24.20	18	0.04	0.15

Q32	21.77	26	<0.01	0.70	30.36	32	<0.01	0.55	13.06	16	<0.01	0.67
Q33	50.51	28	0.06	0.006	52.07	29	0.06	0.005	15.31	16	<0.01	0.50
Q34	17.83	17	0.02	0.40	17.52	18	<0.01	0.49	21.91	13	0.06	0.06
Q35	24.49	24	0.01	0.43	22.84	24	<0.01	0.53	19.21	15	0.04	0.21
Q36	19.54	18	0.02	0.36	18.44	18	0.01	0.43	18.80	16	0.03	0.28
<u>Form4</u>												
Q37	31.70	32	<0.01	0.48	31.94	32	<0.01	0.47	23.90	24	<0.01	0.47
Q38	34.41	30	0.03	0.27	39.84	29	0.05	0.09	27.12	18	0.05	0.08
Q39	19.64	24	<0.01	0.72	20.04	24	<0.01	0.70	14.87	18	<0.01	0.67
Q40	22.42	18	0.04	0.21	24.19	18	0.04	0.15	28.30	19	0.05	0.08
Q41	34.46	33	0.02	0.40	31.84	33	<0.01	0.53	22.17	22	0.01	0.45
Q42	42.01	31	0.05	0.09	35.28	31	0.03	0.27	23.77	19	0.04	0.21
Q43	24.26	26	<0.01	0.56	26.22	27	<0.01	0.51	24.03	19	0.04	0.20
Q44	28.76	29	<0.01	0.48	30.97	29	0.02	0.37	14.85	18	<0.01	0.67
Q45	20.42	16	0.04	0.20	18.72	15	0.04	0.23	12.33	13	<0.01	0.50
Q46	37.19	29	0.04	0.14	29.74	29	0.01	0.43	24.92	18	0.05	0.13
Q47	18.79	21	<0.01	0.60	18.18	23	<0.01	0.75	22.35	18	0.04	0.22
Q48	22.53	26	<0.01	0.66	27.68	26	0.02	0.37	24.77	15	0.06	0.05

Note. S-X^2 = item fit statistic

Appendix 4.9 Person fit

Form 1 (Item 1 to Item 12, N=193)

ID		z.outfit	z.infit	Zh	ID		z.outfit	z.infit	Zh	ID		z.outfit	z.infit	Zh
F1_	59	1.37	1.13	-2.69	F1_	73	0.97	0.11	-2.92	F1_	40	2.57	1.25	-2.98
	73	0.89	0.07	-2.67		40	2.09	0.67	-2.52		80	2.60	2.78	-2.59
	65	0.96	0.87	-2.32		65	0.97	0.88	-2.42		137	2.39	2.48	-2.16
	137	2.00	1.99	-2.16		137	2.10	2.07	-2.34		73	1.90	1.78	-2.04
	40	1.24	0.70	-1.94		80	1.59	2.4	-2.08		97	1.04	1.37	-1.93
	80	1.58	2.31	-1.88		123	1.36	1.26	-1.97		115	0.74	0.69	-1.86
	138	1.34	1.67	-1.85		154	0.93	1.02	-1.76		74	1.88	1.78	-1.78
	5	0.78	0	-1.70		76	0.40	0.32	-1.53		187	1.61	1.58	-1.68
	123	1.34	1.11	-1.68		74	0.78	0.73	-1.43		8	0.97	0.90	-1.38
	41	1.06	0.68	-1.53		162	0.81	0.60	-1.43		47	1.97	1.72	-1.25
	154	0.82	0.95	-1.53		52	0.56	1.01	-1.42		171	0.25	0.04	-1.21
	20	1.26	1.39	-1.50		8	1.16	1.50	-1.38		54	0.47	0.67	-1.2
	97	1.44	1.27	-1.47		187	0.81	1.18	-1.36		41	1.16	0.80	-1.19
	169	0.52	1.69	-1.42		97	0.94	1.26	-1.34		182	1.01	1.56	-1.18
	116	1.69	1.95	-1.39		135	0.62	0.92	-1.32		78	0.80	1.89	-1.12
	66	0.27	0.30	-1.36		111	1.83	1.38	-1.22		76	0.78	1.21	-1.05
	52	0.56	0.98	-1.32		122	1.60	1.66	-1.22		162	0.95	0.64	-1.03
	74	0.74	0.69	-1.27		1	-0.37	-0.24	-1.1		65	1.03	1.20	-0.99
	187	0.76	1.10	-1.19		188	1.94	1.08	-1.1		154	1.11	1.12	-0.97
	111	1.82	1.38	-1.18		115	0.27	0.19	-1.08		64	1.09	1.50	-0.95
	8	1.12	1.40	-1.17		173	0.71	0.97	-1.06		39	1.16	1.09	-0.94
	162	0.67	0.59	-1.14		182	1.24	1.59	-1.03		108	1.39	1.52	-0.92
	122	1.56	1.60	-1.13		138	0.67	1.32	-1.01		68	1.11	0.74	-0.89
	76	0.27	0.11	-1.04		78	0.74	1.59	-0.96		43	0.73	0.19	-0.79
	1	-0.38	-0.27	-1.04		108	1.43	1.92	-0.94		135	0.81	1.33	-0.78
	188	1.90	0.98	-1.02		171	0.41	0.62	-0.93		42	1.71	1.64	-0.74
	135	0.60	0.72	-1.00		116	1.15	1.69	-0.90		178	1.84	1.12	-0.73
	78	1.11	1.64	-0.96		130	0.17	-0.02	-0.81		1	0.94	0.83	-0.71
	108	1.44	1.91	-0.92		191	0.80	0.43	-0.80		175	0.80	0.48	-0.68
	9	1.19	0.55	-0.91		178	1.74	1.03	-0.73		130	0.56	0.91	-0.68
	115	0.22	0.17	-0.89		39	0.28	0.75	-0.70		158	0.83	0.27	-0.64
	182	1.23	1.50	-0.81		175	2.01	1.64	-0.62		123	0.51	0.12	-0.61
	191	0.78	0.42	-0.78		106	0.09	0.12	-0.60		160	0.20	0.25	-0.60
	173	0.66	0.78	-0.69		163	1.25	1.37	-0.57		6	0.11	0.28	-0.49
	75	0.04	0.43	-0.62		118	-0.04	-0.19	-0.57		118	0.33	0.63	-0.48

178	1.50	1.02	-0.62	41	0.53	0.06	-0.57	138	0.45	0.95	-0.46
106	0.10	0.10	-0.60	157	-0.55	-0.64	-0.52	67	0.68	0.99	-0.46
79	0.79	0.85	-0.59	42	0.98	0.81	-0.48	161	0.40	0.44	-0.42
39	0.27	0.70	-0.59	54	-0.27	-0.01	-0.47	36	0.04	-0.20	-0.40
130	0.03	-0.12	-0.52	67	1.00	1.20	-0.46	79	0.56	1.11	-0.37
53	-0.10	-0.43	-0.50	47	0.50	0.90	-0.44	33	0.14	0.02	-0.34
171	0.29	0.39	-0.49	20	0.63	0.89	-0.44	110	0.51	0.42	-0.33
163	1.17	1.30	-0.46	6	0.32	0.98	-0.43	173	0.51	0.91	-0.32
157	-0.63	-0.71	-0.45	83	-0.04	-0.07	-0.39	125	0.08	0.18	-0.30
15	0.41	-0.51	-0.43	161	0.13	-0.06	-0.39	156	-0.05	-0.03	-0.28
146	-0.30	-0.78	-0.42	84	0.17	0.31	-0.38	111	0.14	0.26	-0.28
42	0.88	0.74	-0.37	64	0.92	1.49	-0.37	122	0.18	0.09	-0.28
67	0.95	1.13	-0.35	87	0.03	0.11	-0.33	52	1.00	0.86	-0.25
83	-0.04	-0.08	-0.34	160	0.05	-0.3	-0.31	46	-0.38	-1.4	-0.23
47	0.49	0.83	-0.34	58	0.83	1.67	-0.3	45	0.35	0.67	-0.21
64	0.93	1.47	-0.33	158	0.16	-0.66	-0.29	188	0.43	0.20	-0.20
180	0.54	0.81	-0.33	112	0.32	0.84	-0.28	84	-0.11	-0.06	-0.19
175	1.87	1.54	-0.32	172	0.35	0.05	-0.26	63	0.51	0.47	-0.18
54	-0.32	-0.06	-0.31	183	0.07	0.08	-0.25	116	0.5	0.91	-0.16
161	0.03	-0.08	-0.30	22	-0.83	-0.90	-0.25	58	0.78	1.10	-0.16
34	-0.26	-0.34	-0.29	7	0.52	0.56	-0.25	112	0.1	0.59	-0.14
118	-0.29	-0.40	-0.28	55	1.21	1.19	-0.22	193	0.42	0.52	-0.10
84	0.08	0.24	-0.28	110	0.27	-0.09	-0.21	87	-0.21	-0.4	-0.10
55	1.25	1.21	-0.26	62	0.58	0.59	-0.21	26	-0.33	-0.29	-0.07
62	0.60	0.60	-0.23	79	0.44	0.61	-0.2	94	-0.54	-0.55	-0.05
6	0.29	0.86	-0.22	141	-0.03	0.07	-0.18	95	-0.14	-0.14	-0.04
158	0.09	-0.73	-0.21	68	0.42	0.19	-0.12	90	0.07	0.08	-0.03
141	-0.01	0.09	-0.19	33	-0.10	-0.3	-0.07	83	0.10	0.11	-0.02
22	-0.90	-0.94	-0.19	36	-0.09	-0.36	-0.07	15	0.49	-1.38	-0.01
58	0.81	1.61	-0.19	125	-0.04	0	-0.07	20	0.24	0.40	0.02
112	0.23	0.72	-0.18	26	0.25	0.35	-0.06	179	-0.05	0.29	0.03
183	0	0.02	-0.18	121	-1.26	-1.59	0	59	-0.08	-0.05	0.07
87	-0.06	0.04	-0.15	127	-0.58	-0.37	0.05	44	0.24	0.98	0.10
172	0.26	-0.12	-0.08	169	0.04	0.91	0.08	5	-0.58	-1.96	0.11
38	0.30	0.46	-0.05	156	-0.36	-0.38	0.11	91	-0.18	0.07	0.11
68	0.39	0.19	-0.03	45	-0.03	0.25	0.12	98	0.06	-0.09	0.13
33	-0.15	-0.33	0.02	103	-0.25	-0.84	0.13	114	-0.02	0.18	0.13
7	0.49	0.39	0.04	82	0.26	0.63	0.15	103	0.23	-0.04	0.16
160	-0.20	-0.49	0.04	59	-0.11	-0.33	0.15	9	0.06	-0.15	0.16
81	-0.38	-0.88	0.04	90	-0.05	-0.07	0.16	163	0.50	0.62	0.17

179	-0.09	-0.03	0.05	9	0.37	0.06	0.18	55	-0.12	-0.32	0.17
21	-0.09	-0.46	0.07	69	0.39	-0.05	0.25	82	0.02	-0.35	0.18
125	-0.19	-0.07	0.07	95	-0.70	-1.10	0.27	113	-0.22	-0.81	0.19
121	-1.40	-1.68	0.08	50	0.05	0.27	0.28	50	-0.09	0	0.20
26	0.13	0.24	0.08	179	-0.25	-0.10	0.29	69	0.42	-0.08	0.22
127	-0.59	-0.39	0.12	85	0.01	0.32	0.30	132	-0.42	-1.57	0.22
35	-0.16	-0.08	0.14	92	0.43	1.20	0.32	143	0.21	0.74	0.27
45	-0.06	0.22	0.15	193	0.02	0.23	0.33	106	0.10	0.21	0.28
110	-0.01	-0.31	0.17	98	-0.03	-0.46	0.33	148	-0.08	0.34	0.29
98	0.05	-0.27	0.21	143	0.07	0.36	0.33	136	-0.31	-0.28	0.29
96	-0.58	-0.82	0.27	63	0.07	-0.22	0.33	129	0.27	0.04	0.32
103	-0.45	-1.01	0.30	149	0.34	1.21	0.35	177	0.17	-1.20	0.34
82	0.16	0.44	0.32	44	0.06	0.86	0.35	157	-0.14	0	0.34
36	-0.43	-0.48	0.33	53	-1.09	-1.07	0.37	124	-0.45	-1.43	0.36
50	-0.05	0.18	0.33	4	-0.61	-0.87	0.37	192	-0.68	-0.79	0.38
156	-0.59	-0.48	0.36	91	-0.22	0.44	0.38	183	-0.17	-0.21	0.4
95	-0.83	-1.20	0.36	113	-0.27	-0.86	0.39	75	-0.10	-0.06	0.41
143	-0.01	0.28	0.38	168	-0.88	-1.42	0.4	22	-0.15	-0.01	0.41
4	-0.67	-0.90	0.39	176	0.69	0.92	0.41	35	-0.05	-0.34	0.42
90	-0.43	-0.28	0.4	46	-0.37	-1.43	0.41	92	0.13	0.32	0.43
57	-0.18	0.14	0.42	192	-1.27	-1.67	0.41	169	-0.01	0.70	0.45
63	-0.03	-0.3	0.42	166	0.13	0.62	0.43	11	-0.74	-0.79	0.46
193	-0.14	0.2	0.44	43	-0.71	-0.73	0.43	127	-0.52	-0.63	0.47
192	-1.37	-1.71	0.49	24	-1.26	-1.34	0.45	190	-0.90	-1.35	0.48
168	-0.99	-1.49	0.50	147	-0.50	-0.47	0.47	60	-0.19	-0.02	0.50
44	-0.05	0.73	0.50	35	-0.06	-0.28	0.47	121	-0.33	-0.30	0.50
24	-1.26	-1.31	0.51	136	-0.68	-0.80	0.47	147	-0.17	-0.05	0.52
136	-0.77	-0.87	0.53	29	-0.16	0.10	0.47	85	0.13	0.51	0.53
43	-0.84	-0.81	0.53	129	0.19	-0.09	0.47	166	-0.25	-0.52	0.54
60	-0.74	-0.95	0.53	114	-0.33	-0.27	0.47	104	-0.19	-0.09	0.56
46	-0.48	-1.50	0.54	94	-0.11	0	0.49	191	-0.43	-0.59	0.57
29	-0.29	0.01	0.55	15	0.28	-1.41	0.49	176	-0.02	-0.03	0.59
37	-1.33	-1.09	0.55	60	-0.69	-0.90	0.49	172	-0.34	-0.49	0.59
102	0.05	0.26	0.56	5	-0.81	-1.87	0.51	184	-0.98	-1.56	0.59
86	-0.84	-0.97	0.57	11	-0.76	-0.84	0.52	3	-0.40	-0.14	0.60
56	-0.36	-0.71	0.58	102	0.04	0.27	0.53	168	-0.26	-0.80	0.60
166	-0.06	0.50	0.58	37	-1.31	-1.06	0.53	128	-0.16	-0.08	0.61
91	-0.38	0.34	0.59	119	-0.15	-0.12	0.53	29	0.15	0.54	0.61
94	-0.17	-0.06	0.60	140	-0.28	-0.73	0.53	185	-0.63	-1.48	0.62
11	-0.81	-0.90	0.60	75	-0.28	-0.20	0.59	149	0.01	0.20	0.63

69	-0.26	-0.13	0.61	34	-0.69	-1.10	0.61	49	-0.35	0.10	0.63
114	-0.49	-0.41	0.62	23	-0.89	-0.71	0.62	102	-0.67	-0.83	0.64
176	0.66	0.7	0.62	132	-0.38	-1.07	0.64	14	-0.31	0.06	0.64
92	0.29	0.96	0.65	99	-0.99	-0.89	0.64	189	-1.33	-1.79	0.66
186	0.05	0.25	0.65	18	-0.22	0.13	0.64	186	-0.6	-0.62	0.66
128	-0.46	-0.31	0.66	148	-0.51	-0.21	0.65	117	-0.11	0.51	0.68
147	-0.82	-0.59	0.67	124	-0.45	-1.36	0.65	180	-0.31	-0.06	0.68
23	-0.93	-0.76	0.67	186	0.05	0.26	0.66	131	-0.3	-0.92	0.68
185	-0.73	-1.23	0.68	70	-1.09	-1.47	0.68	96	-0.6	-1.51	0.71
99	-1.05	-0.93	0.68	146	-1.39	-1.80	0.68	140	-0.13	-0.43	0.71
113	-0.80	-1.08	0.73	104	-0.24	-0.44	0.68	99	-0.18	0.06	0.72
70	-1.30	-1.57	0.74	38	-0.36	0.03	0.69	56	-0.37	-0.66	0.74
132	-0.66	-1.04	0.74	57	-0.31	-0.02	0.70	107	-0.43	0.06	0.75
149	0.03	0.89	0.74	180	-0.12	0.26	0.73	23	-0.59	-1.01	0.77
148	-0.69	-0.32	0.77	184	-0.76	-1.07	0.74	119	-0.05	0.03	0.78
12	-0.33	0.10	0.80	86	-0.73	-1.12	0.75	71	-0.89	-1.00	0.80
18	-0.47	-0.01	0.81	96	-0.58	-1.29	0.75	109	-0.55	-0.55	0.80
140	-0.77	-0.96	0.82	12	-0.30	0.12	0.76	93	-0.4	-1.18	0.81
85	-0.78	-0.24	0.82	117	-0.22	0.61	0.80	7	-0.2	-0.14	0.82
105	-0.29	0.19	0.83	128	-0.36	-0.22	0.80	10	-0.75	-0.80	0.85
184	-0.96	-1.19	0.85	164	-0.33	0.42	0.81	34	-0.59	-1.00	0.86
129	-0.34	-0.37	0.88	185	-0.58	-1.36	0.81	142	-0.15	0.09	0.86
124	-0.83	-1.47	0.90	131	-0.31	-0.95	0.82	21	-0.84	-0.84	0.87
117	-0.37	0.49	0.91	177	0.01	-1.25	0.82	89	-1.35	-1.34	0.88
164	-0.47	0.31	0.92	93	-0.38	-1.00	0.83	31	-1.02	-1.06	0.88
31	-0.34	-0.36	0.92	190	-1.02	-1.14	0.86	146	-0.96	-0.91	0.90
177	-0.57	-1.15	0.94	31	-0.27	-0.32	0.86	18	-0.28	-0.11	0.90
119	-1.01	-0.72	0.94	142	-0.18	-0.26	0.90	126	-1.05	-1.29	0.92
104	-0.76	-0.75	0.95	56	-0.44	-0.90	0.90	61	-1.00	-1.01	0.92
10	-0.68	-0.34	0.95	10	-0.60	-0.27	0.93	174	-1.19	-1.57	0.93
190	-1.38	-1.24	1.01	105	-0.18	0.18	0.93	70	-0.68	-0.62	0.93
49	-0.24	0.04	1.04	49	-0.17	0.1	0.95	167	-0.51	-1.15	0.94
19	-0.88	-1.15	1.04	167	-0.53	-0.94	0.98	51	-0.49	-0.19	0.95
3	-0.93	-0.80	1.05	14	-0.79	-0.57	0.99	37	-0.75	-0.75	0.97
61	-0.78	-0.57	1.05	61	-0.74	-0.54	1.02	150	-0.78	-1.16	0.98
14	-0.89	-0.64	1.05	3	-0.92	-0.78	1.03	12	-0.57	-0.34	0.98
150	-1.46	-1.41	1.08	150	-1.43	-1.38	1.04	2	-0.56	-0.58	0.98
167	-1.19	-1.09	1.10	189	-1.27	-1.71	1.06	4	-0.65	-0.63	0.99
133	-1.15	-1.27	1.12	133	-1.09	-1.18	1.07	133	-0.62	-0.77	1.00
131	-0.86	-1.10	1.12	107	-0.77	-0.54	1.09	62	-0.89	-0.86	1.02

89	-0.70	-0.74	1.13	165	-0.58	-0.92	1.10	170	-0.39	-0.95	1.02
107	-0.87	-0.62	1.15	81	-1.26	-1.83	1.10	181	-1.09	-1.1	1.03
142	-0.71	-0.56	1.18	144	-0.34	-1.26	1.10	165	-0.52	-1.05	1.04
189	-1.51	-1.77	1.18	170	-0.38	-0.98	1.12	155	-0.99	-1.19	1.05
109	-0.65	-0.61	1.19	51	-0.52	-0.58	1.13	144	-0.35	-1.44	1.08
93	-1.33	-1.39	1.22	152	-0.43	-1.13	1.14	141	-0.87	-0.83	1.08
144	-1.00	-1.28	1.22	120	-0.28	-0.03	1.14	57	-0.53	-0.78	1.11
2	-1.47	-1.93	1.24	71	-1.05	-0.95	1.15	151	-1.22	-2.09	1.12
51	-0.84	-0.72	1.27	109	-0.63	-0.58	1.19	16	-0.50	0.11	1.12
71	-1.34	-1.01	1.29	134	-0.35	-0.09	1.23	152	-0.40	-1.43	1.16
120	-0.59	-0.20	1.30	2	-1.36	-1.87	1.23	24	-1.28	-1.29	1.17
16	-0.37	0.19	1.31	174	-1.08	-1.66	1.24	81	-1.06	-1.18	1.17
145	-1.53	-1.50	1.33	21	-1.22	-1.17	1.25	105	-0.51	-0.88	1.19
174	-1.32	-1.73	1.34	126	-0.95	-1.34	1.25	159	-2.07	-2.31	1.20
153	-0.76	-0.38	1.35	145	-1.36	-1.43	1.25	38	-1.14	-1.17	1.21
27	-1.51	-1.36	1.36	16	-0.29	0.29	1.26	53	-1.53	-1.46	1.22
165	-1.31	-1.16	1.36	181	-0.94	-1.02	1.30	86	-0.57	-0.95	1.22
139	-1.40	-1.61	1.37	153	-0.72	-0.34	1.34	13	-1.06	-1.69	1.26
181	-1.07	-1.09	1.37	151	-1.03	-1.83	1.34	28	-1.27	-1.68	1.27
170	-0.91	-1.13	1.37	27	-1.44	-1.31	1.34	145	-1.07	-0.56	1.29
100	-2.33	-2.55	1.37	100	-2.36	-2.60	1.36	100	-1.25	-2.06	1.30
126	-1.23	-1.38	1.38	139	-0.75	-1.54	1.39	25	-2.77	-3.95	1.30
134	-0.69	-0.28	1.40	13	-0.95	-1.53	1.40	30	-1.20	-1.28	1.31
32	-0.57	-0.62	1.45	19	-1.09	-1.38	1.45	153	-0.85	-0.55	1.31
151	-1.40	-1.89	1.49	28	-1.11	-1.57	1.49	120	-0.59	-0.83	1.35
152	-1.49	-1.58	1.51	32	-0.61	-0.66	1.50	164	-1.20	-0.91	1.35
13	-1.47	-1.66	1.52	159	-1.63	-1.56	1.52	139	-0.77	-1.77	1.38
17	-0.60	-0.34	1.57	77	-1.02	-1.68	1.53	77	-1.17	-2.00	1.40
28	-1.49	-1.64	1.58	17	-0.58	-0.34	1.58	48	-1.12	-0.95	1.50
159	-1.78	-1.61	1.58	155	-1.33	-1.21	1.60	134	-0.72	-1.10	1.51
155	-1.35	-1.23	1.62	48	-1.03	-1.11	1.63	32	-1.23	-1.25	1.53
30	-1.14	-1.10	1.65	88	-1.12	-1.89	1.63	88	-1.31	-2.38	1.62
77	-1.56	-1.77	1.66	30	-1.14	-1.11	1.64	72	-1.80	-1.81	1.66
48	-1.41	-1.23	1.71	25	-2.21	-2.98	1.66	19	-1.02	-1.11	1.71
25	-2.56	-3.00	1.74	89	-1.06	-1.12	1.77	101	-1.75	-2.46	1.82
88	-1.74	-2.02	1.77	66	-0.97	-1.65	1.84	17	-1.81	-1.67	1.84
72	-1.68	-1.48	1.86	72	-1.50	-1.39	1.86	27	-1.81	-1.67	1.84
101	-1.99	-2.06	1.97	101	-1.48	-1.99	1.91	66	-1.11	-2.00	1.88

Form 2 (Item 13 to Item 24, N=202)

ID		z.outfit	z.infit	Zh	ID		z.outfit	z.infit	Zh	ID		z.outfit	z.infit	Zh
F2_	45	1.51	1.43	-3.20	F2_	62	1.71	1.78	-3.02	F2_	21	2.39	2.28	-2.94
	62	1.73	1.73	-2.96		21	1.59	1.20	-2.45		72	1.89	1.99	-2.28
	22	1.95	1.33	-2.43		20	1.04	1.68	-2.24		132	1.85	1.83	-2.18
	20	0.99	1.55	-1.98		132	1.09	0.95	-2.22		136	1.59	1.21	-1.92
	21	1.20	1.07	-1.92		72	1.48	1.79	-2.03		7	1.46	1.33	-1.89
	132	0.83	0.87	-1.80		85	1.18	1.01	-1.64		20	1.87	2.26	-1.73
	201	1.70	1.77	-1.80		136	1.11	1.21	-1.59		36	0.79	1.98	-1.70
	11	1.01	1.03	-1.75		51	2.08	1.90	-1.50		63	1.34	2.32	-1.54
	37	2.13	2.17	-1.67		151	1.63	1.22	-1.46		85	1.25	0.98	-1.44
	72	1.34	1.58	-1.59		169	0.80	0.98	-1.43		22	1.38	1.05	-1.38
	36	0.78	0.96	-1.49		22	0.89	0.47	-1.43		6	1.19	1.60	-1.29
	128	0.94	1.22	-1.46		36	0.75	0.94	-1.42		101	0.74	0.81	-1.22
	136	1.10	1.14	-1.44		152	1.03	1.43	-1.35		37	2.39	1.57	-1.14
	85	1.00	0.86	-1.34		7	0.55	0.25	-1.13		51	1.25	1.35	-1.09
	169	0.79	0.92	-1.33		75	0.83	0.91	-1.07		75	1.29	1.48	-1.07
	151	1.58	1.13	-1.31		63	1.42	1.77	-1.05		8	0.82	1.30	-1.05
	51	1.81	1.73	-1.20		6	0.76	1.11	-1.03		153	0.76	0.73	-0.99
	104	0.94	0.97	-1.18		87	0.13	-0.03	-0.98		107	1.29	1.24	-0.98
	138	0.42	0.29	-1.11		101	0.57	0.55	-0.96		62	1.17	1.93	-0.97
	63	1.46	1.75	-1.08		23	0.97	1.49	-0.95		145	1.01	0.89	-0.95
	152	0.93	1.19	-0.95		145	0.66	0.83	-0.93		193	0.75	0.56	-0.94
	7	0.50	0.20	-0.94		8	0.13	0.06	-0.90		130	0.71	1.10	-0.92
	61	0.64	0.29	-0.93		104	0.51	0.78	-0.89		61	1.64	0.01	-0.90
	133	0.51	0.42	-0.91		37	0.91	0.86	-0.88		169	0.78	1.37	-0.89
	57	0.66	0.78	-0.88		147	-0.26	-0.72	-0.80		17	1.18	0.62	-0.85
	158	0.47	0.34	-0.88		201	1.12	1.37	-0.80		196	0.99	0.45	-0.81
	156	0.60	0.37	-0.84		167	0.40	0.33	-0.75		87	0.8	0.75	-0.80
	23	1.13	1.38	-0.80		130	0.30	0.39	-0.60		46	0.63	0.73	-0.79
	87	-0.01	-0.20	-0.76		107	0.58	1.33	-0.58		129	0.75	0.58	-0.78
	167	0.42	0.29	-0.76		105	0.60	0.53	-0.58		68	0.44	0.37	-0.78
	8	-0.01	0.01	-0.74		196	0.64	-0.26	-0.55		105	0.78	0.90	-0.77
	75	0.70	0.72	-0.73		118	0.47	0.17	-0.54		64	0.47	0.69	-0.75
	145	0.53	0.70	-0.71		47	0.54	0.77	-0.53		152	0.58	0.80	-0.74
	106	0.84	0.97	-0.70		40	0.20	0.04	-0.51		47	0.37	1.11	-0.68
	24	0.51	0.65	-0.69		68	0.39	0.16	-0.5		168	1.26	0.14	-0.60
	6	0.70	0.92	-0.68		103	0.34	0.95	-0.48		118	0.82	0.76	-0.58
	47	0.60	0.82	-0.53		71	0.58	0.49	-0.47		58	0.37	0.53	-0.56

60	0.63	0.47	-0.52	58	1.39	0.99	-0.46	104	0.87	1.18	-0.56
147	-0.44	-0.77	-0.52	32	-0.09	0.16	-0.45	147	0.56	0.57	-0.52
77	0.33	0.04	-0.48	129	0.07	0.18	-0.44	158	0.37	0.54	-0.52
182	0.77	1.08	-0.48	15	0.3	0.39	-0.44	23	0.46	0.99	-0.52
40	0.2	0.02	-0.47	39	-0.18	-0.38	-0.42	35	0.6	0.59	-0.49
105	0.49	0.43	-0.47	17	0.29	0.13	-0.41	151	0.14	0.26	-0.48
71	0.57	0.45	-0.46	35	0.86	1.02	-0.41	28	0.21	0.42	-0.47
43	0.14	0.25	-0.45	88	0.27	0.45	-0.39	201	0.58	0.31	-0.46
101	0.31	0.29	-0.44	158	0.14	0	-0.35	181	0.51	1.14	-0.46
107	0.52	1.18	-0.41	43	0.11	0.18	-0.35	156	0.53	0.8	-0.44
68	0.32	0.12	-0.4	12	0.53	0.99	-0.32	106	0.25	0.17	-0.42
58	1.33	0.92	-0.4	148	-1.28	-1.3	-0.31	40	0.4	0.38	-0.41
17	0.31	0.12	-0.39	61	0.97	-0.16	-0.31	165	0.18	0.27	-0.38
32	-0.11	0.11	-0.36	181	0.29	0.43	-0.29	172	0.37	0.58	-0.37
12	0.48	0.93	-0.32	193	0.64	0.21	-0.28	2	0.19	0.63	-0.35
181	0.32	0.39	-0.3	159	0.8	1.07	-0.28	192	0.21	0.22	-0.3
39	-0.33	-0.42	-0.28	3	0.89	1.06	-0.28	3	0.32	0.92	-0.29
129	-0.05	0.07	-0.28	168	1.83	0.56	-0.24	103	0.58	0.84	-0.25
83	0.23	0.2	-0.27	97	0.29	0.42	-0.23	49	0.01	0.34	-0.25
103	0.24	0.78	-0.27	192	-0.53	-0.76	-0.21	144	0.11	0.41	-0.24
148	-1.27	-1.27	-0.26	185	0.51	0.04	-0.18	167	0.3	0.85	-0.23
168	1.75	0.69	-0.26	149	0.5	0.21	-0.15	39	0.41	0.78	-0.22
3	0.85	1.01	-0.25	153	0.08	0.06	-0.15	114	0.12	0.03	-0.22
88	0.22	0.38	-0.24	182	0.42	0.95	-0.14	176	-0.04	-0.05	-0.2
35	0.7	0.89	-0.21	106	0.01	-0.01	-0.14	115	-0.04	0.01	-0.18
130	0.12	0.19	-0.2	2	0.59	0.75	-0.12	88	0.25	0.5	-0.16
159	0.69	0.92	-0.19	180	1.05	1.45	-0.12	32	0.4	0.74	-0.13
69	0.3	0.15	-0.18	69	0.28	0.12	-0.12	149	0.55	0.1	-0.12
193	0.52	0.15	-0.15	133	-0.35	-0.23	-0.11	78	0.16	0.1	-0.12
118	0.25	-0.15	-0.14	134	0.01	0.12	-0.09	200	0.8	0.55	-0.11
2	0.59	0.72	-0.13	28	-0.07	0.09	-0.07	119	0.02	-0.25	-0.09
143	0.19	0.09	-0.12	165	-0.03	0.1	-0.07	162	0.3	0.08	-0.08
196	0	-0.34	-0.09	176	-0.2	-0.39	-0.04	133	0.13	0.41	-0.08
192	-0.65	-0.86	-0.05	24	-0.22	0.26	-0.03	19	0.13	0.31	0.03
117	-0.59	-0.69	-0.05	144	-0.34	-0.06	-0.01	108	0.63	-0.1	0.06
131	-0.13	-0.35	-0.04	117	-0.68	-0.95	-0.01	140	0.09	0.04	0.06
67	0.27	-0.6	-0.03	52	-0.13	-0.45	0	190	-0.34	-0.48	0.06
15	-0.17	-0.04	-0.01	156	-0.15	-0.08	0	55	-0.16	0.13	0.07
55	0.42	0.38	0.01	55	0.41	0.41	0.01	5	-0.26	-0.31	0.07
185	0.37	-0.1	0.02	114	0.04	-0.05	0.01	182	0.1	0.62	0.1

92	0.02	-0.16	0.03	108	0.58	0.09	0.01	24	-0.12	0.11	0.12
116	-0.15	-0.1	0.04	46	-0.24	-0.03	0.03	150	0.11	0.26	0.13
46	-0.22	0.05	0.05	150	0.18	0.03	0.04	93	-0.13	-0.14	0.15
49	0.01	-0.1	0.05	49	-0.01	-0.08	0.06	57	-0.11	-0.44	0.18
153	-0.04	-0.04	0.06	5	-0.04	0.14	0.07	9	-0.2	-0.46	0.19
197	-0.04	-0.4	0.06	162	0.29	-0.1	0.08	42	0.12	0.69	0.2
195	-0.11	-0.27	0.06	42	0.19	0.21	0.09	31	-0.13	-0.62	0.22
180	0.95	1.26	0.06	195	-0.23	-0.5	0.11	15	0.14	0.64	0.24
97	0	0.15	0.08	200	-0.31	-0.18	0.14	185	-0.26	-0.53	0.24
42	0.14	0.14	0.1	70	-0.14	-0.15	0.16	43	-0.08	0.17	0.28
134	-0.13	0	0.1	138	-0.53	-0.45	0.16	178	-0.43	-0.47	0.29
33	-0.24	-0.02	0.12	57	0	-0.4	0.25	131	-0.36	-0.26	0.3
200	-0.36	-0.21	0.14	64	0.24	0.48	0.25	53	-0.25	-0.46	0.33
52	-0.34	-0.51	0.16	119	-0.13	-0.15	0.27	159	0.03	0.25	0.33
126	-0.07	-0.33	0.17	54	-0.37	-0.43	0.27	50	-0.22	-0.07	0.34
125	-0.09	-0.62	0.22	56	-0.78	-1.08	0.28	89	-0.63	-0.97	0.36
70	-0.27	-0.24	0.23	91	-0.19	-0.39	0.3	71	-0.11	0.19	0.36
44	-0.18	-0.25	0.25	78	-0.41	-0.46	0.32	163	0.18	0.38	0.36
144	-0.6	-0.27	0.27	140	-0.18	0.05	0.32	179	0	0.16	0.37
176	-0.52	-0.6	0.27	86	-0.58	-0.65	0.33	69	-0.09	0.33	0.38
54	-0.36	-0.44	0.28	172	-0.33	0.34	0.34	54	-0.27	-0.02	0.4
28	-0.38	-0.13	0.29	100	0.09	0.94	0.34	134	-0.07	-0.32	0.41
93	-0.08	-0.37	0.3	93	-0.15	-0.34	0.35	154	-0.02	0.27	0.44
165	-0.35	-0.12	0.32	131	-0.48	-0.66	0.35	100	-0.24	0.08	0.47
64	0.21	0.49	0.33	14	0.04	0.68	0.35	25	0.12	-0.3	0.49
150	-0.07	-0.25	0.33	50	-0.58	-0.65	0.37	60	0.3	-0.35	0.5
5	-0.24	-0.09	0.35	175	-0.62	-0.59	0.37	70	-0.18	0.04	0.52
50	-0.57	-0.66	0.37	33	0.01	-0.08	0.37	11	-0.11	0.09	0.52
175	-0.59	-0.58	0.37	190	-0.57	-0.82	0.37	33	-0.04	-0.43	0.52
91	-0.31	-0.47	0.38	115	-0.43	-0.19	0.39	180	-0.12	0.21	0.54
119	-0.28	-0.29	0.4	18	-0.69	-0.64	0.39	177	-0.53	-0.5	0.55
18	-0.7	-0.65	0.4	191	-0.54	-0.59	0.41	116	-0.1	-0.63	0.56
172	-0.36	0.27	0.41	19	-0.14	0.01	0.41	202	-0.71	-0.77	0.56
86	-0.64	-0.66	0.43	164	-0.1	-0.03	0.44	56	-0.46	-0.47	0.57
140	-0.32	-0.07	0.44	77	-0.46	-0.55	0.45	10	-0.49	-0.55	0.59
122	-0.68	-0.98	0.44	184	-0.05	-0.02	0.49	81	-0.71	-1.02	0.59
149	-0.21	-0.27	0.47	31	-0.05	-0.5	0.5	14	-0.17	0.19	0.59
114	-0.41	-0.32	0.49	11	-0.4	-0.43	0.5	97	-0.37	-0.3	0.6
14	-0.07	0.51	0.49	116	0.04	-0.46	0.51	92	-0.18	-0.6	0.62
164	-0.1	-0.14	0.5	163	0.12	0.1	0.52	12	-0.31	-0.43	0.64

191	-0.67	-0.69	0.51	179	0.02	-0.12	0.52	45	-0.17	-0.45	0.65
56	-1.02	-1.28	0.51	66	-0.98	-1.14	0.52	128	0.01	-0.42	0.67
184	-0.14	-0.12	0.52	143	-0.43	-0.38	0.52	197	-0.97	-1.37	0.68
66	-0.98	-1.18	0.53	128	0.05	-0.09	0.53	109	-0.42	-1.03	0.68
84	-0.35	-0.5	0.54	177	-0.2	0.23	0.53	124	-0.66	-0.45	0.68
78	-0.65	-0.54	0.56	38	-0.68	-0.77	0.56	138	-0.33	-0.23	0.69
108	-0.3	-0.43	0.56	26	-0.07	0.22	0.56	110	-0.59	-0.89	0.7
162	-0.26	-0.37	0.56	92	-0.05	-0.35	0.58	86	-0.37	-0.23	0.7
19	-0.29	-0.1	0.57	120	-1.1	-1.22	0.58	175	-0.07	-1.03	0.7
115	-0.73	-0.39	0.58	53	-0.32	-0.69	0.58	120	-0.52	-0.37	0.71
190	-0.69	-0.92	0.58	60	0.35	-0.08	0.59	122	-0.65	-0.94	0.72
120	-1.1	-1.21	0.59	174	-0.08	0.34	0.62	65	-1.03	-1.05	0.74
59	-0.65	-0.84	0.61	1	-0.49	-0.42	0.63	38	-0.59	-0.62	0.74
177	-0.26	0.16	0.62	124	-0.12	0	0.65	141	-0.65	-0.77	0.74
135	-0.37	0.03	0.62	178	-0.61	-0.37	0.65	26	-0.44	-0.59	0.74
100	-0.13	0.64	0.62	122	-0.73	-1.18	0.65	142	-0.57	-0.57	0.74
1	-0.5	-0.44	0.65	45	-0.09	-0.3	0.66	44	-0.12	-0.61	0.74
124	-0.14	-0.05	0.67	109	-0.33	-0.92	0.72	84	-0.85	-0.65	0.75
38	-0.8	-0.91	0.69	44	-0.2	-0.54	0.75	157	-0.38	-0.05	0.75
111	-0.28	-0.21	0.7	194	-0.17	-0.3	0.78	52	-0.46	-0.51	0.78
178	-0.69	-0.46	0.72	83	-0.65	-0.65	0.8	161	-0.28	-0.49	0.78
174	-0.17	0.23	0.72	135	-0.11	-0.69	0.81	125	-1.1	-1.54	0.81
31	-0.34	-0.57	0.74	171	-0.75	-0.65	0.82	173	-0.73	-0.38	0.81
194	-0.19	-0.34	0.79	29	-0.26	-0.47	0.83	143	-0.41	-0.29	0.82
186	-0.43	-0.73	0.79	146	-1.75	-1.83	0.85	171	-0.47	-1.22	0.82
171	-0.73	-0.64	0.8	112	-0.55	-0.59	0.87	96	-0.66	-0.94	0.83
53	-0.82	-0.95	0.85	41	-0.13	-0.74	0.87	77	-1.1	-1.21	0.83
10	-0.39	-0.25	0.86	110	-0.52	-0.84	0.88	188	-1.21	-1.21	0.87
146	-1.72	-1.8	0.87	198	-0.2	0.34	0.88	195	-0.26	-0.89	0.88
112	-0.59	-0.7	0.88	10	-0.4	-0.23	0.88	135	-0.31	-0.96	0.88
29	-0.35	-0.55	0.88	197	-0.82	-1.28	0.9	90	-0.84	-0.89	0.9
154	-0.66	-0.78	0.89	202	-0.76	-0.91	0.9	146	-0.81	-0.72	0.92
26	-0.4	-0.1	0.92	161	-0.75	-0.95	0.9	18	-0.85	-0.97	0.95
161	-0.79	-0.99	0.93	25	-0.07	-0.79	0.93	41	-0.34	-1.08	0.96
34	-0.27	-0.55	0.94	141	-0.7	-0.63	0.93	184	-0.62	-0.61	0.97
48	-0.97	-1.09	0.97	154	-0.67	-0.78	0.94	112	-0.3	-0.77	0.98
157	-0.63	-0.15	0.97	157	-0.57	-0.07	0.95	83	-0.59	-0.46	0.98
179	-0.82	-0.63	0.97	34	-0.32	-0.63	0.96	34	-0.39	-0.94	0.98
198	-0.37	0.12	0.99	84	-0.82	-0.68	0.97	121	-1.01	-1.6	0.99
202	-0.88	-0.95	1	9	-0.63	-1	0.99	117	-0.54	-0.92	0.99

141	-0.83	-0.67	1	48	-0.96	-1.09	0.99	164	-0.54	-0.92	0.99
163	-0.79	-0.54	1.04	199	-0.2	-0.72	1.03	94	-1.53	-1.65	1.01
65	-1.45	-1.61	1.06	65	-1.3	-1.54	1.03	13	-1.28	-1.47	1.02
4	-1.96	-2.03	1.07	90	-0.74	-0.9	1.04	183	-1.28	-1.47	1.02
137	-0.4	-0.12	1.08	96	-0.77	-0.75	1.05	113	-0.68	-1.29	1.08
155	-0.73	-0.99	1.09	123	-0.14	-0.74	1.05	187	-0.75	-1.07	1.11
109	-1.01	-1.27	1.09	121	-0.77	-1.35	1.07	170	-0.32	-1.04	1.12
110	-1.13	-1.14	1.09	81	-0.52	-0.43	1.08	199	-0.18	-1.03	1.12
82	-1.11	-1.25	1.13	82	-0.97	-1.18	1.08	1	-1	-1.2	1.12
142	-1.27	-1.37	1.15	137	-0.38	-0.09	1.08	123	-0.36	-1.06	1.15
9	-1.06	-1.21	1.15	4	-1.99	-2.05	1.08	67	-0.45	-1.12	1.15
89	-1.2	-1.29	1.17	13	-0.84	-0.92	1.09	174	-1.1	-0.7	1.15
173	-1.06	-0.62	1.17	183	-0.84	-0.92	1.09	102	-1.21	-1.49	1.17
81	-0.67	-0.49	1.17	126	-1.02	-1.03	1.09	126	-1.07	-0.91	1.17
13	-0.95	-1.05	1.2	113	-0.49	-1.09	1.09	59	-1.15	-1.11	1.2
183	-0.95	-1.05	1.2	59	-1.07	-1.26	1.09	198	-1.09	-1.03	1.21
102	-1.62	-1.97	1.21	102	-1.28	-1.8	1.1	191	-1.04	-0.95	1.22
41	-1.39	-1.35	1.25	67	-0.24	-0.96	1.12	137	-1.78	-1.96	1.23
90	-1.09	-1.1	1.26	170	-0.49	-0.9	1.13	194	-1.11	-1.35	1.24
121	-1.35	-1.6	1.27	89	-1.12	-1.24	1.13	139	-1.3	-0.87	1.24
187	-0.77	-0.74	1.3	173	-1.03	-0.57	1.13	189	-1.3	-0.87	1.24
96	-1.28	-1.02	1.32	142	-1.27	-1.37	1.18	79	-1.78	-1.88	1.26
25	-1.08	-1.25	1.37	187	-0.81	-0.79	1.34	29	-1.21	-1.17	1.28
30	-2.32	-2.73	1.38	30	-2.24	-2.71	1.35	4	-0.66	-1.31	1.29
123	-1.34	-1.28	1.44	95	-0.93	-1.56	1.37	148	-1.03	-1.66	1.35
170	-1.32	-1.48	1.45	127	-1.2	-1.33	1.39	155	-1.22	-1.39	1.35
113	-1.34	-1.42	1.45	125	-1.21	-1.49	1.39	66	-0.73	-1.56	1.35
139	-0.93	-0.32	1.46	99	-0.92	-1.5	1.41	99	-1.16	-1.8	1.37
189	-0.93	-0.32	1.46	186	-0.53	-1.25	1.43	95	-1.2	-1.95	1.37
188	-0.65	-0.22	1.46	155	-0.95	-1.19	1.43	48	-1.29	-1.42	1.4
199	-1.45	-1.6	1.48	139	-0.89	-0.22	1.45	111	-2.03	-1.95	1.43
94	-1.53	-1.72	1.51	189	-0.89	-0.22	1.45	91	-1.24	-1.23	1.46
127	-1.66	-1.57	1.54	94	-1.44	-1.67	1.47	160	-1.28	-1.99	1.48
95	-1.77	-1.92	1.58	188	-0.65	-0.18	1.5	98	-1.32	-1.71	1.5
73	-1.06	-0.61	1.61	80	-1.18	-1.16	1.6	186	-0.79	-1.62	1.53
76	-1.69	-1.63	1.64	73	-1.01	-0.5	1.6	127	-1.3	-1.24	1.58
99	-1.66	-1.82	1.64	76	-1.55	-1.49	1.61	73	-1.92	-1.73	1.63
160	-1.34	-1.27	1.65	79	-1.62	-1.77	1.65	166	-1.32	-1.7	1.67
80	-1.41	-1.35	1.66	166	-1.03	-1.39	1.66	30	-1.99	-2.01	1.67
79	-1.73	-1.81	1.69	160	-1.39	-1.32	1.68	80	-1.96	-2.22	1.68

	166	-1.73	-1.68	1.87		111	-1.36	-0.87	1.71		74	-2.33	-2.78	1.74
	98	-1.19	-0.9	1.88		27	-1.27	-2.05	1.73		27	-1.64	-2.65	1.78
	27	-2.23	-2.42	1.91		74	-1.78	-2.19	1.88		82	-2.46	-2.59	1.79
	74	-2.11	-2.32	1.94		98	-1.16	-0.85	1.89		16	-2.5	-2.91	1.81
	16	-2.32	-2.5	2.01		16	-1.97	-2.35	1.96		76	-2.03	-1.89	1.85

Form 3 (Item 25 to Item 36, N=204)

ID		z.outfit	z.infit	Zh	ID		z.outfit	z.infit	Zh	ID		z.outfit	z.infit	Zh
F3_	22	2.90	2.44	-3.91	F3_	63	1.89	2.17	-2.68	F3_	63	2.55	2.32	-3.45
	36	1.94	1.93	-2.88		22	2.02	1.76	-2.55		22	1.75	1.59	-2.27
	204	1.08	0.41	-2.70		92	1.14	0.48	-2.32		72	1.95	1.59	-2.22
	151	2.09	2.63	-2.69		72	1.61	1.61	-2.09		71	2.08	1.51	-2.13
	63	1.83	2.06	-2.49		171	-0.13	-0.10	-1.99		74	1.26	1.09	-2.02
	92	1.08	0.43	-2.21		71	1.31	0.85	-1.86		75	1.55	1.98	-1.97
	20	0.91	1.08	-2.03		75	0.89	0.89	-1.80		73	1.72	1.09	-1.61
	111	1.02	1.31	-2.01		195	0.81	0.76	-1.68		92	1.20	1.48	-1.58
	171	-0.13	-0.08	-1.91		156	1.92	2.54	-1.61		70	3.10	1.47	-1.47
	72	1.55	1.51	-1.84		73	1.57	1.37	-1.59		156	1.23	1.80	-1.30
	71	1.30	0.74	-1.72		204	0.11	-0.38	-1.33		97	1.13	1.38	-1.26
	195	0.73	0.70	-1.63		74	1.11	1.20	-1.33		19	1.48	1.60	-1.19
	75	0.72	0.75	-1.52		19	1.94	1.82	-1.31		26	0.95	1.20	-1.13
	182	1.10	1.14	-1.52		150	-0.29	-0.38	-1.07		52	0.97	1.05	-1.12
	177	0.87	0.46	-1.52		101	1.42	1.00	-1.01		182	0.76	0.69	-1.11
	156	1.81	2.47	-1.48		45	0.17	0.11	-0.98		195	0.92	1.14	-1.07
	191	1.14	1.09	-1.36		182	0.61	0.87	-0.98		46	0.96	1.10	-1.02
	19	1.84	1.71	-1.13		157	0.75	0.13	-0.97		191	0.85	1.26	-0.95
	74	1.05	1.09	-1.07		87	0.12	0.28	-0.97		9	0.83	0.61	-0.87
	150	-0.29	-0.41	-1.03		191	0.33	0.34	-0.94		157	0.81	0.88	-0.85
	157	0.76	0.12	-0.98		97	0.80	1.31	-0.89		96	0.66	-0.01	-0.85
	45	0.15	0.09	-0.95		166	0.68	0.76	-0.87		194	0.83	0.65	-0.84
	166	0.67	0.76	-0.85		151	1.03	1.86	-0.86		44	1.11	1.80	-0.84
	159	0.85	0.75	-0.84		130	1.10	1.35	-0.85		34	1.24	1.77	-0.83
	68	0.61	0.67	-0.81		154	1.98	1.79	-0.82		107	0.64	1.19	-0.81
	73	0.99	1.10	-0.80		194	1.04	0.89	-0.81		62	0.83	0.77	-0.73
	194	0.99	0.85	-0.77		52	0.48	0.48	-0.78		130	0.32	0.63	-0.67
	130	1.02	1.31	-0.76		85	0.40	0.56	-0.75		85	0.42	0.54	-0.64
	87	0.04	0.20	-0.75		46	0.79	1.27	-0.72		111	0.70	0.76	-0.61
	154	1.79	1.75	-0.66		43	0.15	0.78	-0.69		151	0.82	1.19	-0.59
	101	1.15	0.86	-0.63		27	1.30	1.44	-0.64		204	0.77	0.75	-0.57

108	0.76	0.95	-0.62	177	0.01	-0.28	-0.61	177	0.26	0.37	-0.54
97	0.57	1.13	-0.58	172	-0.21	-0.32	-0.6	101	0.42	0.38	-0.53
79	0.19	-0.42	-0.57	102	-0.31	-0.7	-0.6	43	0.68	1.02	-0.51
34	1.09	1.49	-0.56	34	1.1	1.51	-0.58	125	0.33	-0.31	-0.5
27	1.31	1.35	-0.55	70	1.12	0.92	-0.58	147	0.58	1.28	-0.48
17	0.04	-0.32	-0.55	26	0.37	0.63	-0.56	122	0.75	0.19	-0.43
70	1.11	0.89	-0.54	17	0.03	-0.35	-0.55	154	0.79	0.78	-0.41
42	0.48	0.29	-0.51	47	0.29	0.48	-0.55	162	0.15	0.44	-0.39
102	-0.37	-0.75	-0.5	9	0.58	0.13	-0.53	144	0.46	0.73	-0.39
26	0.36	0.55	-0.49	96	0	-0.75	-0.53	175	0.81	1.28	-0.37
43	0.09	0.65	-0.48	164	-0.15	0.28	-0.52	87	0.63	1.04	-0.36
175	-0.19	0.43	-0.47	175	-0.18	0.46	-0.5	202	0.38	0.79	-0.35
1	-0.92	-1.05	-0.46	1	-0.88	-1.01	-0.49	132	0.4	0.36	-0.34
164	-0.19	0.24	-0.44	111	0.26	0.51	-0.49	155	0.07	0.37	-0.28
168	0.55	0.68	-0.43	140	0.37	1.34	-0.36	3	0.3	1.09	-0.28
85	0.29	0.35	-0.41	137	1.07	1.07	-0.35	115	0.12	0.24	-0.25
142	0.25	-0.11	-0.41	107	0.33	-0.35	-0.35	139	0.44	0.16	-0.23
52	0.16	0.34	-0.41	6	1.36	0.79	-0.33	127	-0.11	-0.43	-0.23
163	-0.35	-0.69	-0.4	44	1.14	1.43	-0.33	143	0.23	0.13	-0.21
47	0.11	0.34	-0.39	120	0.03	-0.03	-0.28	47	0.41	0.58	-0.21
106	0.88	1.61	-0.38	183	0.25	0.41	-0.28	137	0.27	0.31	-0.2
107	0.33	-0.35	-0.35	66	0.93	0.29	-0.28	126	0.3	0.4	-0.19
137	1.04	1.07	-0.33	202	0.21	0.19	-0.25	119	0.32	-0.47	-0.13
44	1.13	1.39	-0.31	20	0.66	1.1	-0.22	106	0.2	0.85	-0.12
6	1.3	0.73	-0.31	3	0.4	0.74	-0.22	24	-0.22	-0.26	-0.11
96	-0.14	-0.81	-0.3	62	0.11	0.5	-0.21	150	-0.28	-0.22	-0.09
120	0.02	-0.06	-0.29	57	0.02	-0.08	-0.2	140	0.29	1	-0.07
183	0.27	0.39	-0.27	133	0.67	1.42	-0.2	20	0.29	0.37	-0.06
46	0.5	1.1	-0.27	51	0.23	0.85	-0.19	133	0.33	0.85	-0.06
172	-0.41	-0.53	-0.21	125	0.23	-0.69	-0.19	66	-0.08	-0.23	-0.06
57	-0.01	-0.12	-0.18	178	0.36	0.02	-0.19	60	0.56	1.13	-0.04
202	0.16	0.17	-0.18	56	0.52	-0.08	-0.15	8	0.18	0.57	-0.01
66	0.75	0.23	-0.17	95	-0.16	-0.38	-0.1	102	0.36	0.59	-0.01
178	0.39	-0.05	-0.16	106	0.34	1.29	-0.09	27	0.31	0.37	-0.01
3	0.31	0.68	-0.14	192	0.65	0.4	-0.08	25	0.32	-0.1	0
62	0.07	0.45	-0.12	189	-0.35	-0.09	-0.06	172	0.04	0.34	0
95	-0.13	-0.33	-0.11	197	-0.31	0.09	-0.05	169	-0.1	0	0.02
51	0.17	0.77	-0.1	139	0.14	0.49	-0.05	89	0.14	0.58	0.03
9	0.25	0.01	-0.09	145	0.21	1.06	-0.04	45	-0.21	-0.53	0.04
56	0.45	-0.14	-0.04	131	-1.21	-1.48	-0.03	30	-0.39	-0.66	0.05

187	0.03	-0.35	-0.03	60	-0.17	0.14	0	29	-0.19	-0.45	0.07
192	0.56	0.34	-0.01	36	0.52	0.61	0.02	141	0.65	0.06	0.07
133	0.59	1.25	-0.01	24	-0.3	-0.38	0.05	59	-0.05	0.56	0.07
197	-0.34	0.02	0.01	58	0.41	0.62	0.06	112	-0.08	0.11	0.07
140	0.07	0.97	0.04	116	-0.22	-0.13	0.06	121	-0.05	0.23	0.09
139	0.11	0.42	0.05	136	-0.09	0.01	0.07	167	0.22	0.75	0.09
60	-0.22	0.09	0.06	148	-0.91	-1.44	0.08	93	-0.22	-0.12	0.09
125	-0.18	-0.85	0.07	190	0.53	0.01	0.08	51	0.33	0.83	0.1
131	-1.2	-1.47	0.1	167	-0.13	0.8	0.09	168	-0.2	0.02	0.11
141	0.64	0.51	0.1	8	-0.2	-0.06	0.09	14	0.08	0.32	0.12
58	0.36	0.61	0.11	162	-0.08	-0.2	0.1	190	-0.71	-0.8	0.16
136	-0.11	-0.01	0.11	141	0.64	0.49	0.1	201	0.08	0.58	0.17
162	-0.1	-0.2	0.12	127	-0.18	-0.99	0.11	6	-0.02	0.23	0.18
145	0.14	0.88	0.13	144	0.57	0.65	0.11	136	0.07	0.02	0.18
84	-0.53	-0.58	0.13	119	0.16	-0.58	0.12	145	0.1	0.53	0.2
144	0.55	0.64	0.13	168	0.01	0.42	0.15	105	-0.36	-0.8	0.21
158	-0.36	-0.61	0.15	152	-0.48	-0.14	0.18	36	-0.22	-0.14	0.21
190	0.41	-0.03	0.16	89	0.39	0.56	0.19	86	0	-0.13	0.22
189	-0.46	-0.26	0.17	99	0.65	0.34	0.2	166	-0.18	-0.36	0.22
116	-0.44	-0.31	0.18	86	0.22	0.61	0.2	152	0.22	0	0.24
148	-1	-1.43	0.22	118	0.33	1	0.2	197	0.09	0.12	0.24
155	0.04	0.14	0.25	155	0.08	0.18	0.21	99	-0.34	-0.65	0.26
167	-0.21	0.64	0.25	169	-0.23	-0.04	0.22	64	-0.48	-0.17	0.29
99	0.6	0.28	0.25	29	-0.27	-0.81	0.25	118	-0.04	0.26	0.3
40	0.23	0.09	0.26	115	0.08	0.11	0.28	110	-0.32	-0.08	0.31
86	0.18	0.55	0.26	64	-0.96	-0.54	0.29	187	-0.41	-0.49	0.32
152	-0.54	-0.22	0.26	126	-0.52	-0.12	0.29	134	-0.17	-0.83	0.32
127	-0.33	-1.03	0.27	147	-0.55	-0.25	0.3	164	0	0.49	0.37
89	0.31	0.52	0.27	122	0.17	-0.03	0.31	120	-0.17	0.09	0.37
115	0.07	0.09	0.29	134	-0.33	-1.21	0.31	1	-0.4	-0.24	0.37
169	-0.37	-0.13	0.29	13	-0.67	-0.97	0.34	98	-0.6	-0.67	0.38
24	-0.67	-0.61	0.31	163	-1.08	-1.36	0.34	33	0.03	-0.11	0.39
126	-0.54	-0.13	0.33	105	-0.27	-0.82	0.36	159	-0.16	-0.14	0.39
122	0.16	-0.05	0.35	128	-0.98	-1.5	0.37	57	0.14	-0.42	0.4
134	-0.28	-1.22	0.35	179	0.13	0.05	0.37	142	-0.3	-0.29	0.41
64	-1	-0.61	0.37	193	-0.59	-0.62	0.37	58	-0.44	-0.36	0.41
143	-0.49	-0.56	0.37	143	-0.5	-0.56	0.38	53	-0.36	-0.31	0.42
147	-0.61	-0.29	0.37	200	-0.34	0.18	0.39	200	-0.52	-0.42	0.42
29	-0.52	-0.93	0.43	25	-0.15	-0.02	0.4	13	-0.48	-0.44	0.43
179	0.03	0.05	0.44	14	-0.36	0.11	0.41	192	-0.22	0.3	0.44

153	-0.09	-0.11	0.44	108	0.08	0.62	0.42	170	-0.69	-0.87	0.45
119	-0.27	-0.69	0.45	132	-0.23	-0.25	0.44	68	-0.07	0.07	0.45
132	-0.24	-0.26	0.45	33	0.05	-0.1	0.45	108	-0.09	0.26	0.47
59	-0.42	-0.42	0.46	159	-0.12	-0.27	0.45	131	-0.58	-0.67	0.49
25	-0.19	-0.11	0.46	153	-0.11	-0.14	0.46	2	-0.59	-0.48	0.49
54	-1.08	-1.21	0.47	93	-0.03	0.46	0.46	128	-0.72	-1.16	0.49
13	-0.79	-1.03	0.48	201	-0.36	-0.58	0.46	42	-0.01	-0.25	0.51
201	-0.4	-0.59	0.49	121	-0.47	-0.41	0.47	10	-0.4	-0.75	0.52
128	-1.07	-1.58	0.49	54	-1.07	-1.19	0.47	40	-0.13	-0.62	0.53
112	-0.2	-0.14	0.51	112	-0.14	-0.1	0.47	193	-0.52	-0.41	0.53
121	-0.52	-0.42	0.52	187	-0.4	-0.62	0.47	183	-0.23	-0.24	0.54
200	-0.4	0.06	0.53	110	-0.3	-0.05	0.48	146	-0.47	-0.55	0.56
93	-0.09	0.38	0.55	59	-0.44	-0.45	0.48	135	-0.28	-0.16	0.58
88	-0.44	-0.73	0.56	30	-0.61	-1.14	0.49	116	-0.51	-0.59	0.59
14	-0.46	-0.03	0.57	10	-0.25	-0.6	0.51	181	-1.03	-1.43	0.61
30	-0.69	-1.18	0.57	170	-0.73	-1.45	0.51	82	-0.75	-0.35	0.61
110	-0.4	-0.15	0.6	146	-0.68	-1.29	0.52	84	-0.66	-1.11	0.64
4	-0.53	-0.74	0.6	40	-0.04	-0.53	0.52	180	-0.83	-0.98	0.67
146	-0.83	-1.36	0.6	68	-0.2	-0.25	0.59	189	-0.37	-0.01	0.7
170	-0.86	-1.49	0.64	4	-0.52	-0.7	0.6	174	-1.01	-1.26	0.7
8	-0.86	-0.5	0.64	12	-0.23	0.03	0.6	5	-0.34	-0.37	0.72
105	-0.6	-1.05	0.66	142	-0.43	-0.6	0.6	16	-0.9	-0.82	0.72
174	-1.15	-1.56	0.68	11	-0.33	0.39	0.6	55	-0.65	-0.53	0.73
123	-0.42	-0.51	0.69	84	-0.51	-0.83	0.61	11	-0.36	0.1	0.74
196	-0.38	-0.37	0.7	104	-0.03	0.43	0.63	35	-0.45	-0.48	0.74
118	0.03	0.6	0.7	135	-0.18	-0.38	0.65	49	-0.44	-0.07	0.74
193	-1.1	-0.98	0.71	98	-0.87	-0.76	0.67	100	-0.57	-0.57	0.75
11	-0.45	0.21	0.72	174	-1.08	-1.54	0.7	54	-0.61	-0.51	0.75
12	-0.36	-0.15	0.75	196	-0.38	-0.36	0.7	163	-0.72	-0.72	0.76
53	-0.05	0.03	0.76	180	-0.49	-0.5	0.76	67	-0.17	-0.49	0.79
50	-1.26	-1.11	0.77	49	-0.07	0.75	0.76	12	-0.8	-0.85	0.79
98	-0.99	-0.85	0.8	53	-0.04	0.04	0.77	113	-0.75	-1.16	0.79
104	-0.18	0.24	0.81	50	-1.3	-1.15	0.78	165	-1.2	-1.38	0.79
10	-0.85	-0.94	0.81	55	-0.97	-0.46	0.79	4	-0.6	-0.75	0.82
180	-0.56	-0.58	0.82	67	-0.21	-0.28	0.82	48	-0.87	-0.46	0.84
35	-0.04	0.09	0.83	35	-0.05	0.07	0.83	94	-1.28	-1.7	0.85
203	-0.62	-0.79	0.84	161	-0.27	-0.07	0.83	56	-1.4	-1.44	0.86
55	-1.01	-0.51	0.84	117	-0.53	-0.81	0.84	161	-0.51	-0.79	0.86
5	-0.84	-0.85	0.85	5	-0.86	-0.87	0.86	117	-0.73	-0.97	0.86
135	-0.41	-0.6	0.88	203	-0.66	-0.8	0.86	15	-0.67	-0.57	0.87

2	-0.92	-1.02	0.91	123	-0.42	-0.56	0.87	123	-0.51	-0.7	0.87
31	-0.53	-0.02	0.94	184	-0.45	-0.78	0.88	17	-0.64	-0.82	0.92
161	-0.41	-0.24	0.95	181	-0.79	-1.59	0.88	95	-0.64	-0.82	0.92
49	-0.32	0.47	0.98	2	-0.89	-0.99	0.89	129	-0.64	-0.82	0.92
100	-0.73	-0.9	1	100	-0.61	-0.77	0.91	171	-0.64	-0.82	0.92
181	-0.98	-1.64	1.01	15	-0.61	-0.57	0.91	178	-0.64	-0.82	0.92
77	-0.68	-0.99	1.01	16	-0.95	-0.93	0.92	114	-1.08	-0.87	0.92
117	-0.99	-1.06	1.02	31	-0.39	0.11	0.95	184	-0.62	-1.09	0.92
82	-0.74	-0.63	1.03	82	-0.69	-0.59	0.96	104	-0.59	-0.45	0.94
39	-1.04	-1.06	1.03	39	-0.95	-0.93	0.97	31	-0.77	-0.66	0.95
16	-1.07	-1.02	1.03	158	-1.21	-1.17	0.98	186	-0.79	-1.14	0.95
23	-0.74	-0.69	1.04	42	-0.6	-0.52	1	173	-0.75	-0.56	0.95
61	-1.01	-1.08	1.12	186	-0.65	-0.98	1.01	50	-0.82	-0.86	0.96
33	-0.89	-0.59	1.13	61	-0.61	-0.88	1.02	199	-1.4	-1.65	0.97
165	-1.21	-1.64	1.14	90	-0.61	-0.97	1.05	38	-0.79	-0.8	0.98
15	-0.9	-0.75	1.14	23	-0.77	-0.78	1.05	185	-1.06	-1.12	0.99
94	-1.45	-1.29	1.18	185	-0.97	-1.03	1.06	23	-0.62	-1.02	1
184	-1.36	-1.28	1.21	88	-0.68	-0.95	1.12	149	-0.62	-1.02	1
186	-1.04	-1.19	1.22	165	-1.18	-1.61	1.12	153	-0.62	-1.02	1
90	-0.96	-1.17	1.22	109	-1.06	-0.99	1.14	203	-0.62	-1.02	1
114	-1.04	-1.03	1.25	114	-0.92	-0.98	1.16	41	-1	-1.13	1.01
185	-1.36	-1.25	1.25	94	-1.49	-1.28	1.18	18	-0.98	-0.95	1.02
18	-0.48	0.1	1.26	199	-1.09	-1.77	1.23	90	-0.78	-1.19	1.02
38	-0.78	-0.63	1.26	78	-0.8	-0.64	1.23	61	-0.8	-1.01	1.03
109	-1.35	-1.18	1.27	124	-1.19	-2	1.25	148	-0.67	-0.3	1.04
91	-1.57	-1.81	1.3	18	-0.46	0.14	1.26	124	-1.49	-2.05	1.04
199	-1.27	-1.83	1.3	38	-0.76	-0.59	1.26	65	-1.47	-1.53	1.1
188	-1.44	-0.92	1.31	80	-0.85	-1.12	1.27	32	-0.78	-0.98	1.1
103	-1.49	-1.98	1.31	188	-1.42	-0.89	1.29	78	-0.83	-0.54	1.15
124	-1.4	-2.08	1.33	91	-1.52	-1.79	1.29	196	-1.58	-1.79	1.15
32	-1.38	-1.52	1.34	79	-1.63	-2.04	1.3	88	-0.91	-1.27	1.2
78	-1.02	-0.78	1.34	103	-1.16	-1.82	1.3	103	-1.51	-1.93	1.22
149	-1.51	-1.78	1.41	32	-1.42	-1.55	1.37	158	-1.02	-0.95	1.23
129	-1.27	-1.74	1.41	129	-1.26	-1.72	1.4	39	-1.28	-1.44	1.25
113	-1.22	-1.52	1.43	149	-1.5	-1.75	1.41	80	-1.05	-1.36	1.25
80	-1.21	-1.28	1.43	48	-1.51	-1.39	1.45	179	-1.13	-0.96	1.26
48	-1.47	-1.39	1.43	81	-0.87	-0.08	1.45	21	-1.79	-2.2	1.29
7	-1.3	-1.38	1.44	21	-1.54	-1.99	1.45	198	-2.56	-2.75	1.3
81	-0.91	-0.16	1.46	7	-1.3	-1.38	1.45	138	-1.54	-1.38	1.31
21	-1.66	-2.04	1.49	113	-1.26	-1.54	1.45	109	-1.04	-0.85	1.37

65	-1.39	-1.12	1.50	83	-0.81	-1.21	1.46	160	-1.81	-1.47	1.39
67	-1.35	-0.93	1.53	138	-1.29	-1.27	1.48	7	-1.39	-1.58	1.41
138	-1.44	-1.33	1.53	65	-1.43	-1.11	1.52	188	-1.46	-0.95	1.44
173	-1.85	-1.53	1.55	173	-1.78	-1.49	1.54	91	-1.83	-1.89	1.45
28	-1.75	-1.76	1.56	28	-1.76	-1.75	1.57	79	-1.73	-1.88	1.46
76	-2.17	-2.57	1.60	76	-2.04	-2.50	1.60	83	-0.99	-1.66	1.49
160	-1.67	-1.35	1.64	160	-1.60	-1.31	1.62	28	-1.66	-1.84	1.58
176	-1.67	-1.47	1.65	41	-1.46	-1.38	1.63	81	-1.55	-1.10	1.66
41	-1.51	-1.43	1.65	176	-1.68	-1.46	1.66	37	-1.89	-2.15	1.70
69	-1.34	-1.12	1.66	69	-1.30	-1.06	1.67	76	-2.17	-2.29	1.72
198	-2.11	-2.31	1.69	198	-2.10	-2.30	1.69	77	-1.57	-1.85	1.73
83	-1.67	-1.58	1.76	77	-1.29	-1.48	1.71	69	-2.29	-2.41	1.81
37	-2.10	-2.07	2.03	37	-2.10	-2.06	2.04	176	-1.96	-1.80	1.81

Form 4 (Item 37 to Item 48; N=180)

ID	z.outfit	z.infit	Zh	ID	z.outfit	z.infit	Zh	ID	z.outfit	z.infit	Zh
F4_ 3	1.62	1.14	-2.19	F4_ 3	1.64	1.18	-2.30	F4_ 118	2.16	2.07	-2.51
134	1.76	1.63	-1.99	63	2.17	2.02	-2.11	63	1.98	1.95	-2.15
118	1.20	1.06	-1.94	118	1.24	1.12	-2.06	70	1.98	2.10	-2.03
142	1.39	1.16	-1.87	142	1.42	1.23	-2.01	50	1.82	1.60	-1.98
63	2.11	1.84	-1.75	72	0.82	0.77	-1.84	61	1.13	1.36	-1.38
50	1.23	0.43	-1.73	50	1.25	0.45	-1.81	142	1.25	1.11	-1.36
70	1.32	1.64	-1.66	132	0.86	0.94	-1.77	75	1.32	1.47	-1.27
174	0.97	1.5	-1.56	70	1.34	1.68	-1.74	132	1.21	1.34	-1.23
132	0.83	0.88	-1.55	95	1.35	1.67	-1.54	60	0.72	0.49	-1.09
95	1.29	1.62	-1.43	140	1.80	1.83	-1.43	76	0.86	0.93	-1.01
140	1.74	1.75	-1.34	81	2.57	2.36	-1.29	74	1.41	0.94	-1.00
124	0.05	0.23	-1.19	104	0.76	0.32	-1.24	17	1.16	1.39	-0.96
104	0.76	0.3	-1.15	61	0.60	0.73	-1.23	81	1.75	1.55	-0.95
42	0.42	0.99	-1.15	42	0.43	1.03	-1.18	95	1.01	1.21	-0.94
91	0.85	0.85	-1.10	91	0.89	0.92	-1.18	104	0.98	0.75	-0.85
29	0.93	0.39	-0.97	44	0.30	0.21	-1.13	144	1.05	1.18	-0.85
81	2.17	2.23	-0.95	5	2.59	2.07	-0.98	44	0.94	1.03	-0.79
5	2.55	2.02	-0.93	29	0.95	0.39	-0.98	140	0.98	1.04	-0.77
61	0.47	0.54	-0.90	30	0.06	-0.06	-0.92	121	0.48	0.22	-0.76
44	0.21	0.14	-0.89	144	0.40	0.34	-0.90	154	1.16	0.90	-0.70
30	0.04	-0.07	-0.87	75	0.66	0.76	-0.88	3	0.89	0.62	-0.69
180	0.39	0.37	-0.85	180	0.40	0.40	-0.85	30	0.48	0.80	-0.69
144	0.38	0.32	-0.84	137	-0.32	-0.17	-0.84	91	0.86	0.90	-0.68
75	0.64	0.73	-0.84	8	0.79	0.69	-0.81	28	0.64	0.67	-0.67

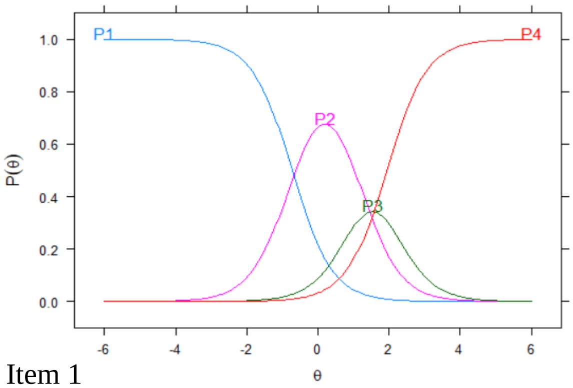
137	-0.31	-0.17	-0.79	28	1.34	1.53	-0.8	179	0.79	0.93	-0.64
74	1.23	0.98	-0.73	179	0.99	1.09	-0.8	69	0.6	0.71	-0.63
88	-0.2	-0.42	-0.72	74	1.26	1.03	-0.78	72	0.66	0.88	-0.59
179	0.93	1	-0.67	88	-0.15	-0.39	-0.76	42	0.93	1.12	-0.57
72	-0.17	0.01	-0.63	76	0.39	0.38	-0.63	169	0.26	0.06	-0.57
28	1.26	1.43	-0.61	65	0.21	0.4	-0.62	33	0.7	1.34	-0.56
8	0.72	0.57	-0.6	161	-0.32	-0.4	-0.57	161	0.66	0.67	-0.48
77	0.49	0.35	-0.59	158	0.59	0.2	-0.49	130	0.58	0.42	-0.47
58	0.56	0.41	-0.55	60	0.58	0.19	-0.47	167	0.6	0.59	-0.44
161	-0.35	-0.43	-0.53	92	0.8	1.07	-0.46	178	0.31	0.18	-0.43
126	0.24	0.15	-0.52	134	1.21	1.05	-0.45	139	0.22	0.41	-0.4
6	0.31	-0.14	-0.5	154	0.53	0.21	-0.41	65	0.77	1.08	-0.38
94	0.48	-0.01	-0.47	165	0.8	1.15	-0.37	8	0.48	0.48	-0.37
9	0.27	0.19	-0.45	117	-0.8	-0.5	-0.36	79	0.19	-0.05	-0.37
158	0.56	0.15	-0.43	86	0.37	0.13	-0.36	37	0.43	0.54	-0.36
60	0.55	0.16	-0.43	167	-0.27	-0.68	-0.36	39	0.2	0.44	-0.36
165	0.8	1.14	-0.39	141	0.83	1.04	-0.35	90	0.28	0.03	-0.34
92	0.74	1	-0.38	150	0.16	0.08	-0.35	5	0.26	0.28	-0.23
65	0.06	0.21	-0.35	172	0.85	0.81	-0.35	29	0.18	0.31	-0.23
76	0.25	0.23	-0.35	127	1.63	2.11	-0.28	158	0.39	0.09	-0.21
117	-0.82	-0.51	-0.35	39	0.14	0.29	-0.28	25	0.23	0.19	-0.2
141	0.84	1.02	-0.34	139	0.09	0.24	-0.27	123	0.43	0.62	-0.2
86	0.35	0.14	-0.34	107	-0.54	-0.51	-0.26	1	0.32	0.31	-0.17
172	0.84	0.79	-0.32	32	0.13	0.11	-0.25	85	0.33	0.48	-0.15
18	0.38	0.24	-0.31	153	1.21	0.54	-0.25	107	0.31	0.54	-0.15
167	-0.27	-0.68	-0.31	37	-0.11	-0.24	-0.23	58	0.03	0.22	-0.14
150	0.14	0.07	-0.3	109	-0.08	-0.55	-0.23	134	0.52	0.61	-0.14
127	1.59	2.04	-0.24	135	0.98	1.08	-0.21	46	0.19	0.43	-0.12
109	-0.09	-0.57	-0.22	90	-0.02	-0.37	-0.2	180	0.33	0.36	-0.11
32	0.09	0.06	-0.21	1	0.42	0.47	-0.17	87	0.09	-0.11	-0.1
153	1.19	0.5	-0.21	100	-0.03	0.14	-0.17	150	0.02	0.06	-0.1
107	-0.57	-0.52	-0.2	21	-0.23	0.15	-0.16	164	0.2	0.23	-0.1
154	0.29	0.14	-0.18	121	0.21	0.08	-0.14	137	0.5	0.41	-0.09
101	0	-0.04	-0.17	85	0.37	0.52	-0.1	135	0.24	0.34	-0.08
100	-0.04	0.13	-0.16	17	0.62	0.39	-0.1	12	0.12	0.28	-0.07
37	-0.17	-0.29	-0.16	58	-0.03	0.14	-0.09	172	0.13	0.09	-0.01
135	0.91	1.05	-0.14	33	0.58	1.14	-0.09	163	0.18	0.18	0.01
1	0.4	0.42	-0.13	130	0.4	0.34	-0.08	166	0.21	0.48	0.01
53	0.06	0.18	-0.09	178	0.6	0.45	-0.07	21	0.25	0.75	0.02
121	0.17	0.05	-0.07	123	0.1	0.06	-0.07	47	-0.06	-0.26	0.02

90	-0.06	-0.43	-0.07	11	0.3	0.57	-0.06	96	0.19	0.28	0.04
130	0.37	0.29	-0.04	25	0.49	0.53	-0.05	9	-0.07	0.01	0.06
17	0.56	0.35	-0.04	169	-0.25	-0.27	-0.05	92	0.29	0.36	0.06
11	0.27	0.55	-0.02	166	0.5	1.03	-0.03	128	-0.17	0.11	0.07
85	0.32	0.47	-0.02	41	-0.09	0.02	-0.02	127	0.08	0.39	0.08
41	-0.1	0.01	-0.01	151	0.61	0.65	0.05	149	0.24	0.11	0.11
178	0.58	0.41	0	12	0.17	-0.05	0.05	151	0.08	0.23	0.11
7	0.14	0.35	0	56	0.47	0.66	0.05	57	0.01	0.62	0.12
25	0.47	0.49	0	22	0.01	-0.11	0.06	153	-0.05	-0.24	0.12
110	0.05	0.26	0.02	69	-0.06	-0.28	0.07	165	0.51	0.51	0.12
123	0.01	-0.02	0.02	164	-0.47	-0.57	0.07	67	-0.43	-0.74	0.13
69	-0.04	-0.25	0.04	136	1.3	1.82	0.07	88	0	0	0.14
56	0.45	0.64	0.06	176	-0.39	-0.39	0.09	22	-0.01	-0.2	0.15
12	0.15	-0.07	0.09	128	-0.13	0.11	0.09	136	-0.1	0.13	0.15
169	-0.29	-0.3	0.1	79	-0.09	-0.5	0.1	48	-0.03	-0.18	0.17
33	0.41	1.03	0.1	83	0.23	-0.01	0.11	77	-0.16	0.08	0.23
21	-0.39	0.04	0.1	43	0.63	0.52	0.14	89	-0.27	-0.42	0.24
176	-0.4	-0.4	0.1	133	0.03	0.04	0.18	32	0.06	0.17	0.26
22	-0.01	-0.12	0.11	170	-1.94	-1.93	0.2	11	-0.05	0.16	0.27
136	1.26	1.76	0.11	77	-0.07	0.11	0.2	171	-0.43	-0.61	0.30
164	-0.53	-0.62	0.12	19	-0.18	-0.95	0.23	2	-0.03	-0.67	0.31
39	-0.11	-0.06	0.12	9	-0.19	-0.16	0.25	51	-0.44	-0.64	0.33
151	0.51	0.57	0.13	112	-0.76	-0.55	0.29	133	0.12	0.28	0.34
139	-0.1	-0.01	0.14	163	-0.21	-0.44	0.3	141	-0.22	0.15	0.34
166	0.38	0.95	0.14	71	-0.13	0.04	0.32	49	-0.12	0.03	0.35
131	-0.16	0.16	0.14	87	-0.33	-0.43	0.35	94	-0.35	-0.34	0.35
170	-1.89	-1.87	0.19	96	0.02	0.19	0.36	56	-0.36	-0.29	0.36
43	0.61	0.47	0.2	174	-0.05	0.29	0.36	93	-0.49	-0.43	0.36
19	-0.16	-0.93	0.21	124	-0.79	-0.69	0.37	175	-0.18	-0.16	0.37
79	-0.13	-0.55	0.22	2	-0.41	-0.4	0.37	71	-0.26	-0.1	0.38
83	0.18	-0.08	0.26	26	-0.62	-0.61	0.38	148	-0.52	-0.81	0.41
163	-0.25	-0.47	0.31	156	0.22	0.08	0.39	119	-0.33	-0.33	0.43
112	-0.76	-0.55	0.32	20	-0.03	0.59	0.39	145	-0.63	-0.66	0.45
96	-0.03	0.13	0.36	16	0.08	0.57	0.4	16	-0.1	0.07	0.46
20	-0.04	0.57	0.37	89	-0.3	0.04	0.41	34	-0.33	-0.47	0.46
26	-0.65	-0.63	0.39	94	-0.45	-0.47	0.43	177	-0.36	-0.57	0.46
89	-0.26	0.07	0.39	119	-0.23	-0.22	0.43	174	-0.22	-0.14	0.47
2	-0.45	-0.43	0.41	175	0.31	0.68	0.46	111	-0.35	-0.25	0.5
156	0.18	0.03	0.41	146	-1.09	-0.95	0.46	131	-0.26	-0.16	0.5
133	-0.16	-0.25	0.42	168	-0.18	0.26	0.47	20	0.02	0.15	0.52

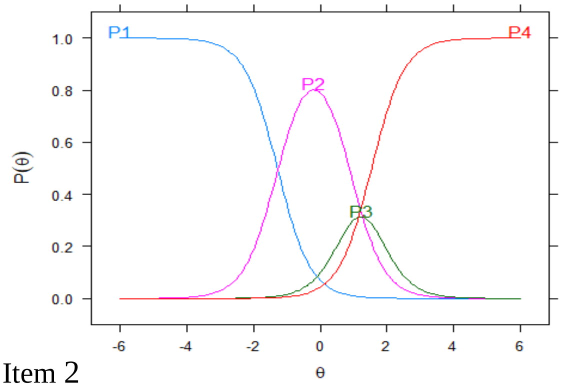
146	-1.06	-0.92	0.43	173	0.03	-0.04	0.48	124	-0.33	-0.09	0.53
87	-0.38	-0.49	0.43	145	0.35	0.27	0.5	155	-0.58	-0.72	0.54
16	-0.03	0.5	0.47	46	-0.74	-0.49	0.51	168	-0.26	-0.14	0.54
175	0.29	0.67	0.47	6	-0.69	-0.52	0.52	41	-0.16	0.33	0.55
168	-0.24	0.2	0.5	55	0.14	0.68	0.52	170	-0.39	-0.36	0.56
173	0	-0.07	0.51	48	-0.39	-0.89	0.52	83	-0.74	-0.92	0.57
46	-0.72	-0.45	0.51	131	-0.34	-0.26	0.53	117	-0.25	0.32	0.57
55	0.12	0.63	0.53	171	-0.62	-1.01	0.55	126	-0.65	-0.84	0.57
97	-0.44	-0.07	0.54	49	-0.56	-0.09	0.55	6	-0.31	-0.18	0.59
145	0.32	0.23	0.56	147	-0.72	-0.57	0.58	115	-0.82	-0.93	0.6
128	-0.44	-0.2	0.56	148	-0.38	-0.84	0.6	86	-0.27	0.04	0.61
49	-0.63	-0.16	0.59	155	-0.28	-0.2	0.64	147	-0.45	-0.37	0.61
48	-0.45	-0.94	0.6	98	-1.67	-2.08	0.66	43	-0.69	-0.96	0.62
171	-0.63	-1.03	0.6	93	-1.18	-1.3	0.68	105	-0.76	-1.08	0.63
147	-0.84	-0.68	0.63	126	-0.62	-0.79	0.71	146	-0.42	-0.43	0.64
98	-1.73	-2.08	0.67	106	-0.47	-0.26	0.72	26	-0.31	-0.58	0.69
93	-1.17	-1.27	0.68	54	-0.59	-0.48	0.72	68	-0.66	-0.64	0.69
155	-0.33	-0.27	0.69	152	-0.3	0.2	0.73	156	-0.44	-0.61	0.69
148	-0.44	-0.88	0.72	51	-0.17	-0.22	0.74	84	-0.45	-0.59	0.7
51	-0.17	-0.23	0.73	101	-0.59	-0.55	0.75	120	-0.66	-0.58	0.71
10	0.24	0.6	0.76	103	-0.05	-0.22	0.76	100	-0.44	-0.08	0.72
54	-0.65	-0.54	0.76	10	0.24	0.61	0.76	54	-0.65	-0.72	0.73
152	-0.4	0.14	0.76	120	-0.17	0.09	0.77	106	-0.51	-0.25	0.73
103	-0.07	-0.23	0.79	84	-0.32	-0.59	0.77	4	-0.58	-0.85	0.78
120	-0.22	0.05	0.8	159	-0.55	-0.51	0.78	159	-0.48	-0.42	0.78
119	-0.65	-0.65	0.81	27	-0.91	-0.69	0.8	55	-0.81	-0.52	0.79
84	-0.46	-0.69	0.82	149	-0.65	-0.65	0.8	101	-0.53	-0.45	0.79
149	-0.7	-0.7	0.83	35	-1.32	-1.42	0.81	138	-0.82	-0.76	0.8
125	-0.48	-0.41	0.84	125	-0.48	-0.37	0.81	152	-0.48	-0.36	0.81
71	-0.72	-0.56	0.84	162	-0.27	0.07	0.83	59	-0.7	-0.7	0.83
67	-0.44	-0.41	0.86	122	-0.88	-0.96	0.83	103	-0.86	-1.18	0.84
35	-1.39	-1.46	0.88	67	-0.43	-0.4	0.85	24	-1.08	-1.21	0.85
122	-0.96	-1.04	0.89	138	-0.73	-0.72	0.87	112	-0.7	-0.67	0.85
159	-0.72	-0.71	0.9	115	-1.47	-1.4	0.89	122	-0.72	-0.84	0.86
111	-0.65	-0.46	0.9	18	-0.62	-0.72	0.9	18	-0.82	-0.92	0.9
27	-0.97	-0.74	0.91	111	-0.63	-0.45	0.92	27	-0.87	-0.91	0.9
115	-1.46	-1.4	0.91	105	-1.19	-1.62	0.93	35	-0.97	-1.1	0.91
47	-0.6	-0.54	0.91	47	-0.61	-0.56	0.93	10	-1.13	-0.99	0.95
162	-0.35	-0.04	0.91	24	-1.2	-1.45	0.96	176	-0.56	-0.71	0.95
105	-1.2	-1.62	0.94	108	0.2	-0.05	0.99	162	-0.8	-0.81	0.96

177	-0.17	-0.44	0.96	177	-0.19	-0.46	0.99	19	-0.87	-0.94	0.98
108	0.19	-0.05	0.98	68	-0.89	-1.29	1.00	64	-1.33	-1.66	0.99
102	-1.00	-0.98	0.99	160	-0.39	0.07	1.01	173	-0.86	-0.95	1.00
138	-0.93	-0.92	0.99	57	-1.38	-1.09	1.01	53	-1.19	-1.30	1.01
68	-0.90	-1.29	1.00	102	-1.02	-1.02	1.02	116	-0.80	-0.90	1.01
24	-1.21	-1.48	1.01	59	-1.02	-0.96	1.03	109	-0.96	-0.94	1.05
57	-1.38	-1.09	1.01	7	-0.58	-0.31	1.05	78	-1.27	-1.52	1.06
160	-0.51	-0.06	1.04	36	-2.32	-2.8	1.06	102	-1.22	-1.27	1.09
59	-1.04	-0.97	1.05	116	-0.69	-0.77	1.07	31	-1.03	-1.10	1.10
36	-2.30	-2.75	1.07	53	-0.82	-0.68	1.10	45	-1.03	-1.10	1.10
106	-0.95	-0.83	1.10	113	-0.97	-0.94	1.12	7	-1.20	-1.30	1.15
45	-0.67	-0.48	1.15	45	-0.65	-0.45	1.13	113	-1.24	-1.15	1.16
113	-1.16	-1.12	1.18	97	-0.78	-0.82	1.16	125	-1.27	-1.20	1.16
114	-1.20	-1.27	1.20	78	-0.97	-1.35	1.18	38	-0.98	-1.08	1.17
78	-1.00	-1.38	1.21	114	-1.20	-1.26	1.21	160	-0.92	-0.75	1.17
34	-1.06	-0.77	1.23	34	-1.01	-0.71	1.23	97	-1.06	-1.10	1.22
116	-1.03	-1.09	1.26	80	-1.08	-1.12	1.29	114	-1.01	-1.37	1.24
82	-1.36	-1.55	1.36	110	-0.85	-0.93	1.30	143	-1.25	-1.20	1.24
157	-1.35	-1.55	1.36	157	-1.24	-1.44	1.31	82	-1.51	-1.70	1.25
64	-1.12	-1.73	1.37	82	-1.26	-1.47	1.31	40	-1.81	-1.92	1.32
129	-2.16	-2.33	1.38	64	-1.10	-1.71	1.35	13	-1.68	-1.72	1.33
4	-1.14	-1.13	1.38	129	-2.17	-2.36	1.37	98	-1.12	-1.01	1.33
66	-2.95	-3.29	1.39	66	-2.99	-3.36	1.38	110	-1.12	-1.26	1.34
143	-2.03	-1.94	1.47	52	-1.10	-1.26	1.41	157	-1.65	-1.78	1.34
31	-2.49	-2.55	1.49	4	-1.18	-1.16	1.42	80	-1.46	-1.58	1.37
38	-0.91	-0.76	1.50	31	-2.50	-2.55	1.49	108	-1.56	-1.74	1.37
13	-0.70	-0.39	1.51	38	-0.93	-0.80	1.50	14	-1.79	-1.98	1.38
80	-1.43	-1.35	1.54	13	-0.69	-0.36	1.52	52	-1.36	-1.63	1.42
14	-1.45	-1.76	1.54	143	-2.08	-2.02	1.54	129	-1.71	-1.80	1.43
40	-1.67	-1.89	1.62	14	-1.41	-1.74	1.55	36	-1.80	-1.89	1.51
99	-1.58	-1.34	1.67	40	-1.63	-1.87	1.61	73	-2.24	-2.29	1.57
73	-1.94	-1.98	1.71	99	-1.55	-1.33	1.67	15	-2.38	-2.37	1.62
52	-1.98	-1.99	1.78	73	-1.91	-1.95	1.69	99	-2.05	-2.01	1.72
15	-1.94	-2.04	1.84	15	-1.94	-2.05	1.85	23	-2.74	-2.87	1.84
23	-2.17	-2.35	1.96	23	-2.13	-2.35	1.99	62	-2.74	-2.87	1.84
62	-2.17	-2.35	1.96	62	-2.13	-2.35	1.99	66	-2.55	-2.58	1.90

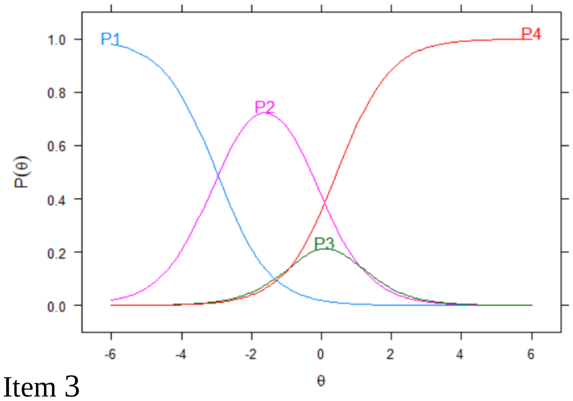
Appendix 4.10 The Category Characteristic Curves of the 4-Category EI Accuracy Scores



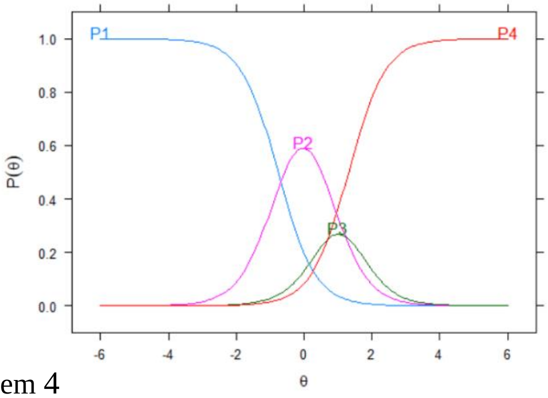
Item 1



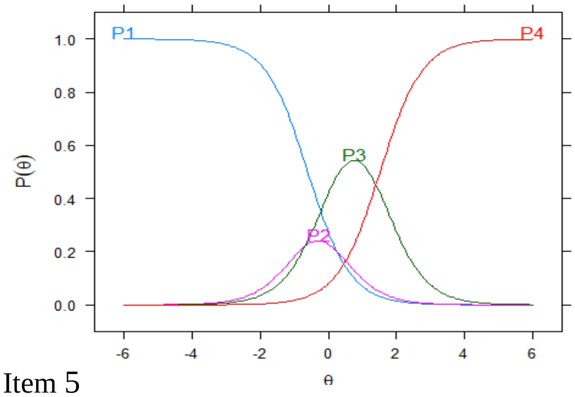
Item 2



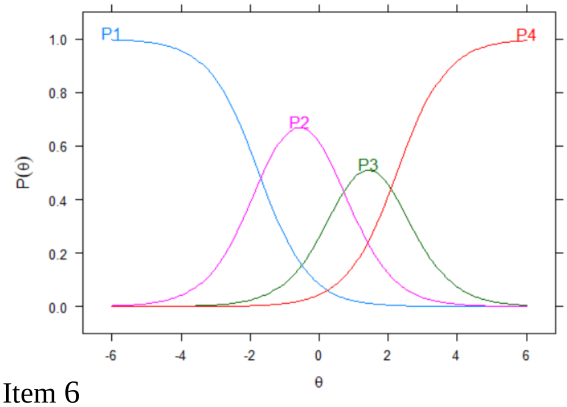
Item 3



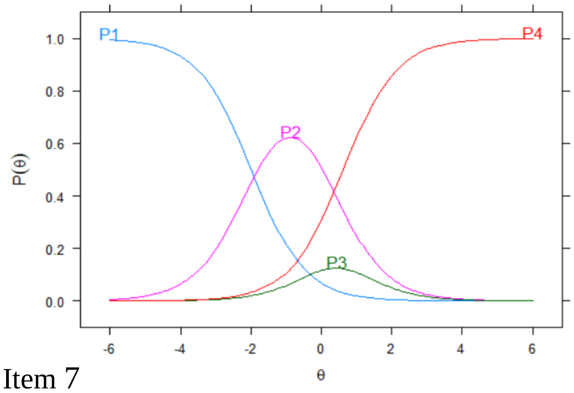
Item 4



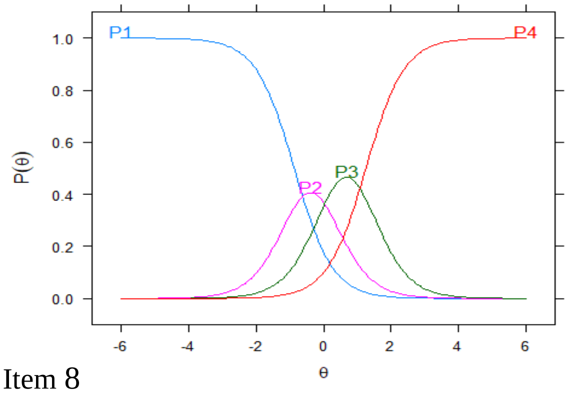
Item 5



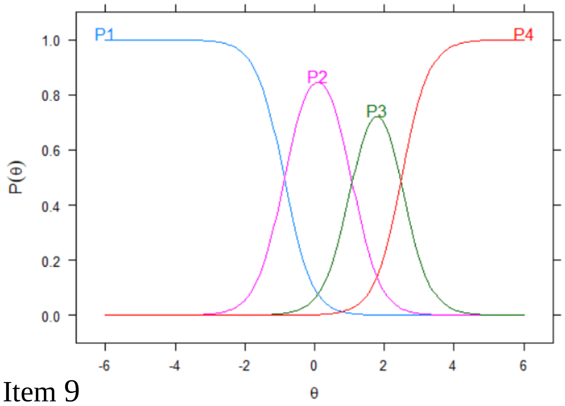
Item 6



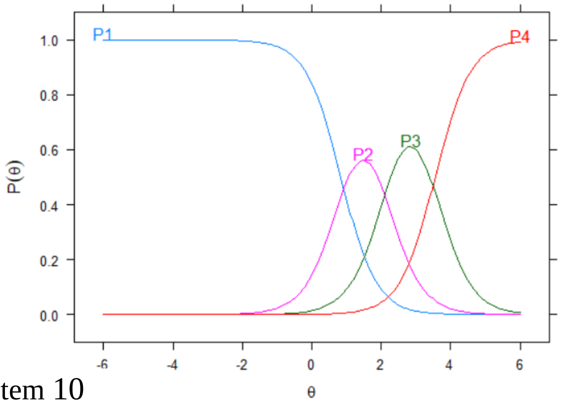
Item 7



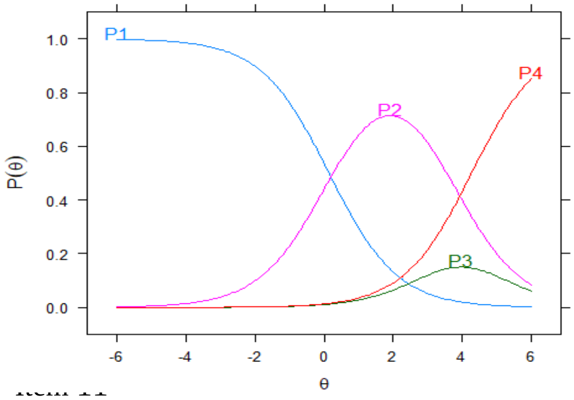
Item 8



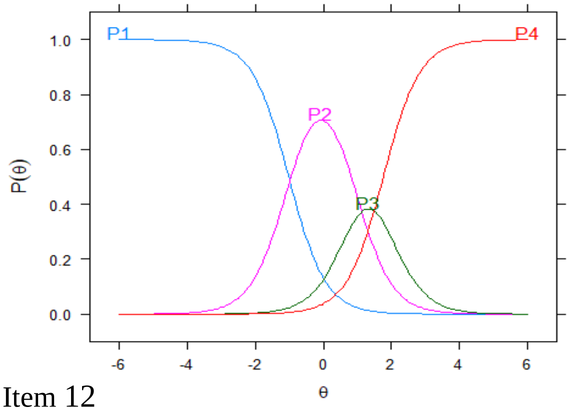
Item 9



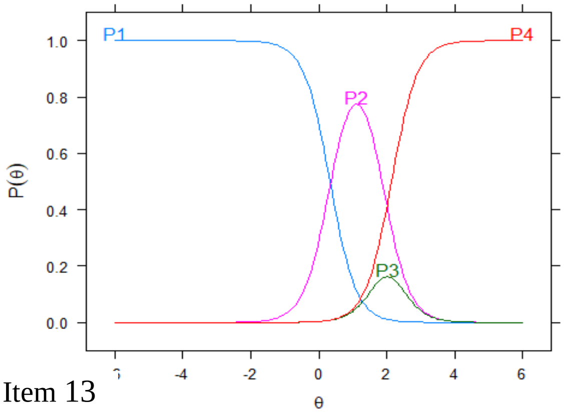
Item 10



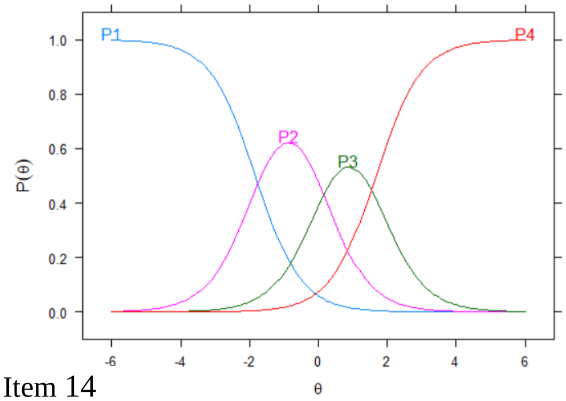
Item 11



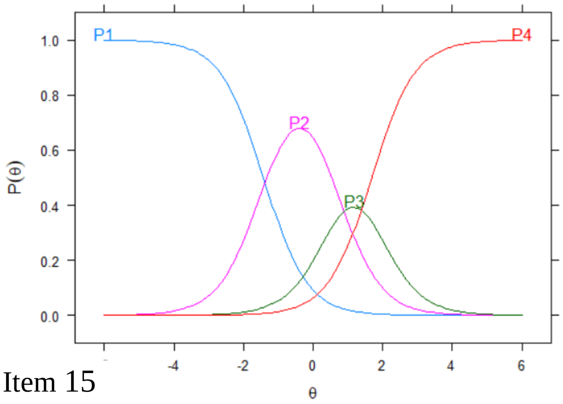
Item 12



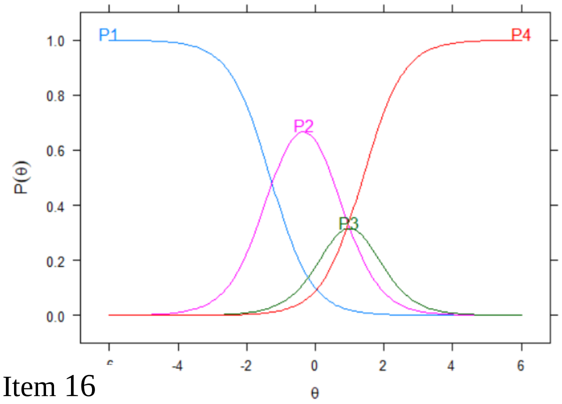
Item 13



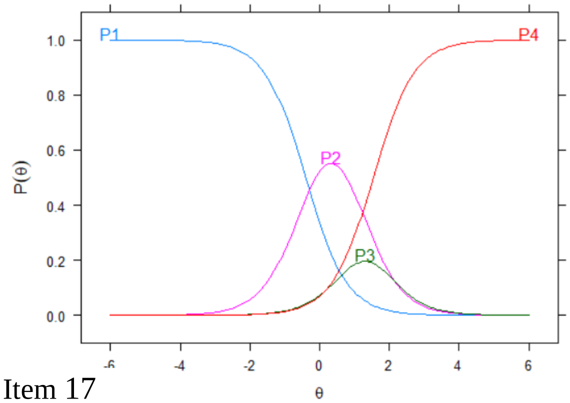
Item 14



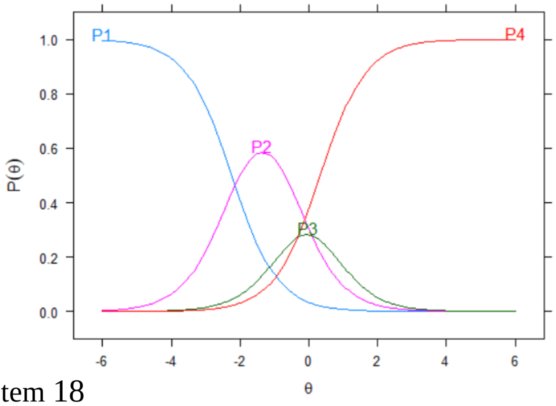
Item 15



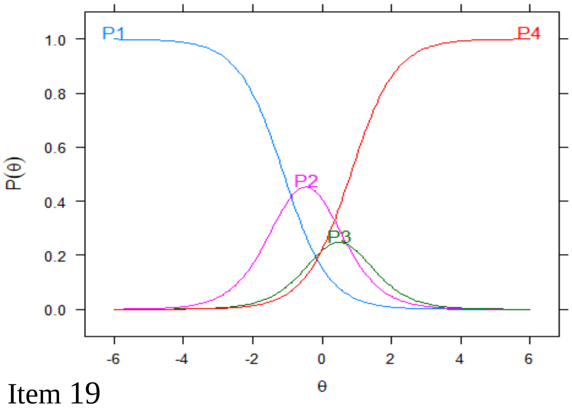
Item 16



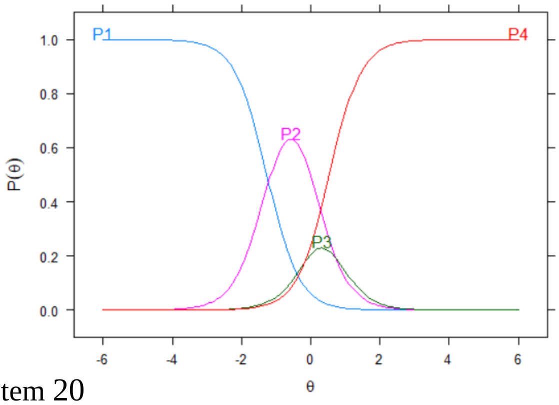
Item 17



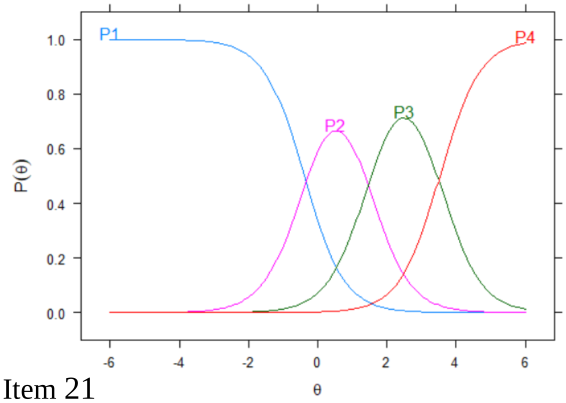
Item 18



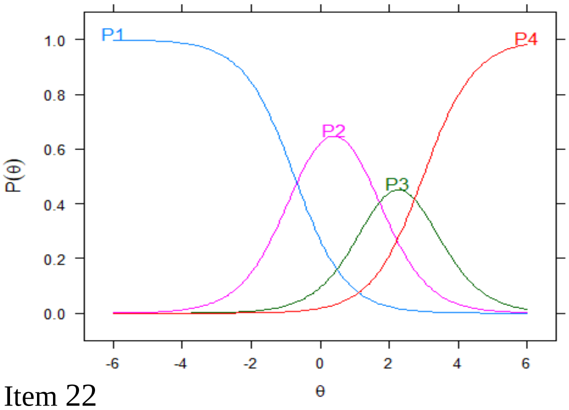
Item 19



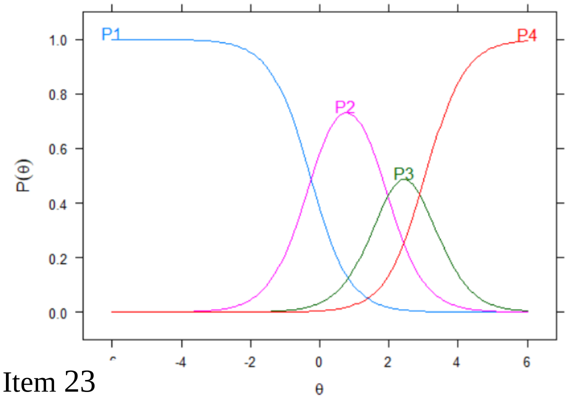
Item 20



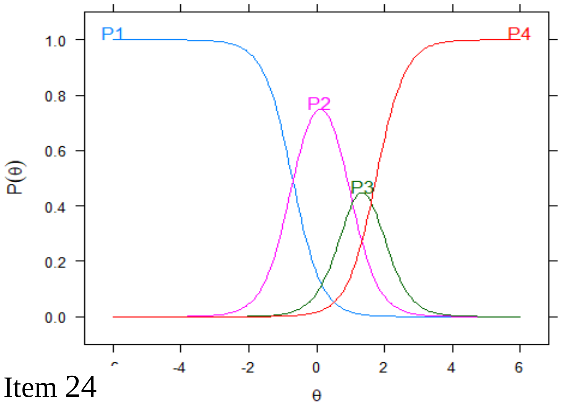
Item 21



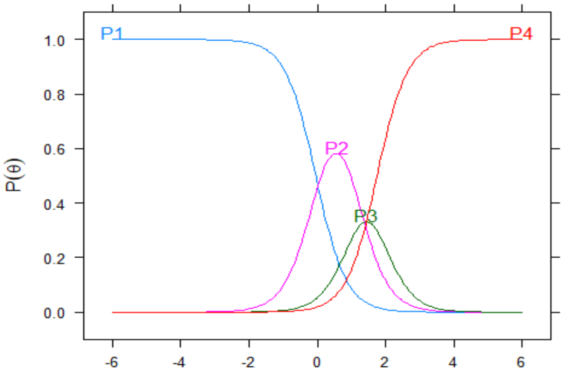
Item 22



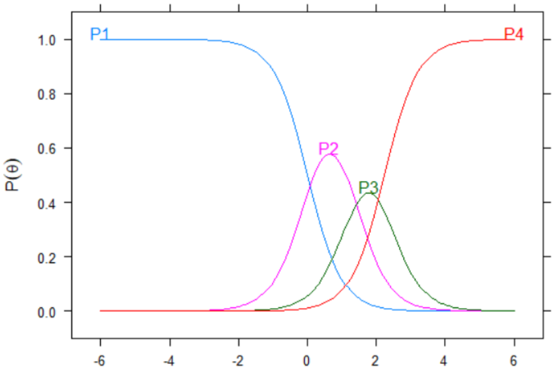
Item 23



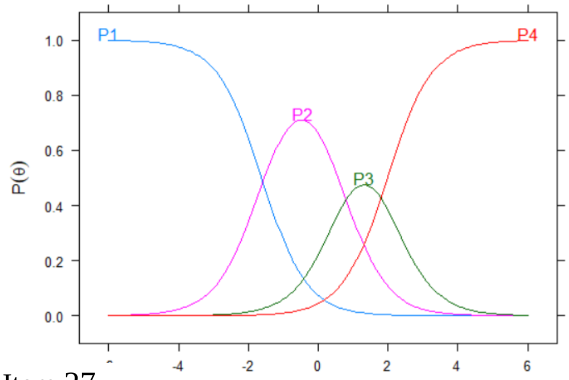
Item 24



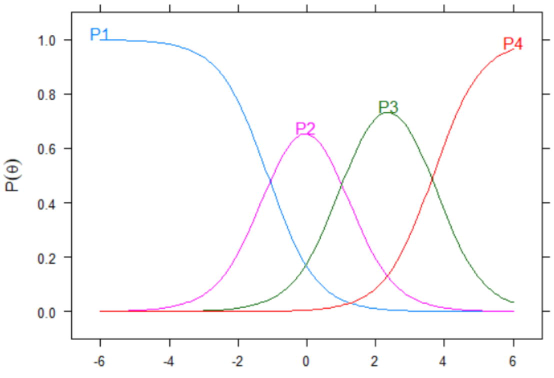
Item 25



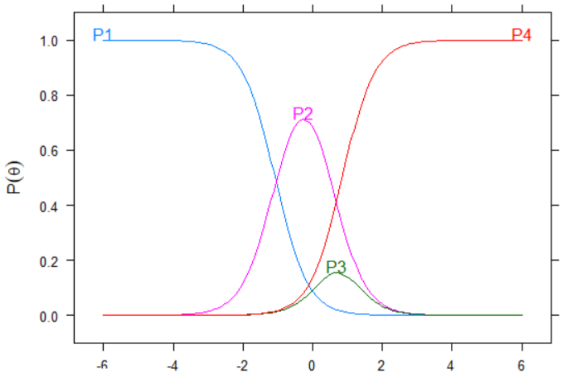
Item 26



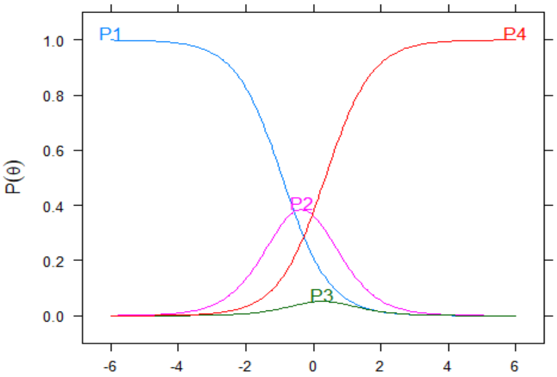
Item 27



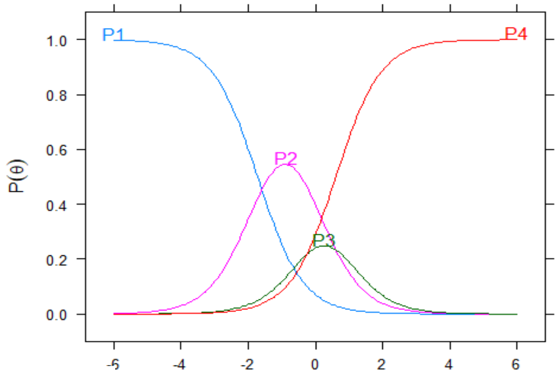
Item 28



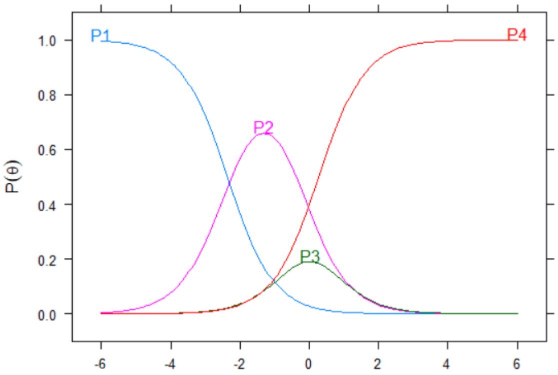
Item 29



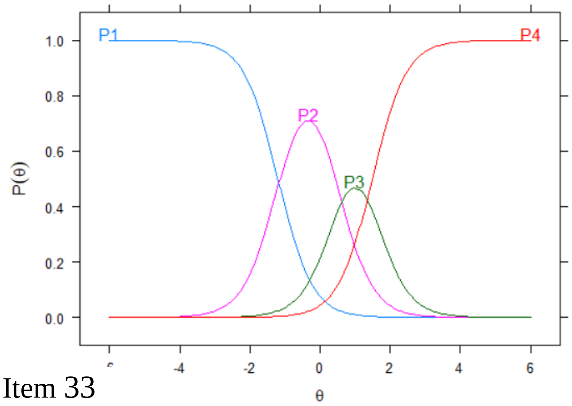
Item 30



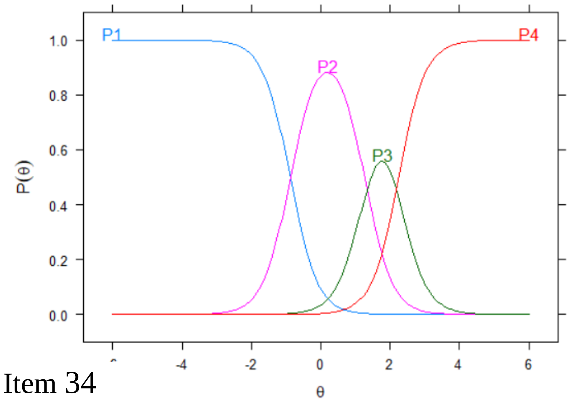
Item 31



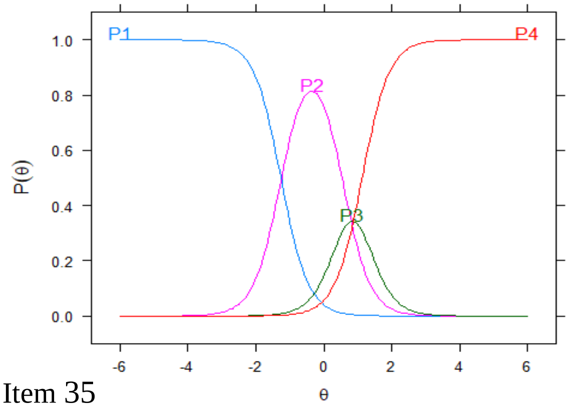
Item 32



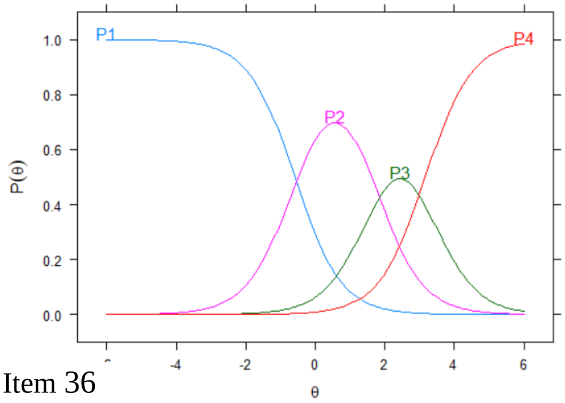
Item 33



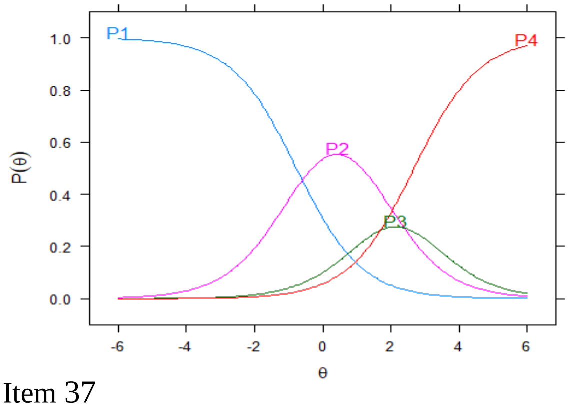
Item 34



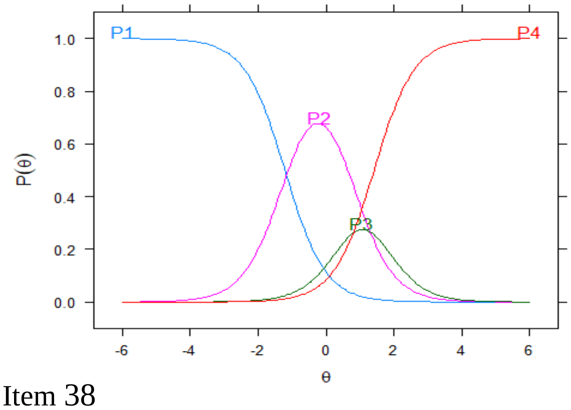
Item 35



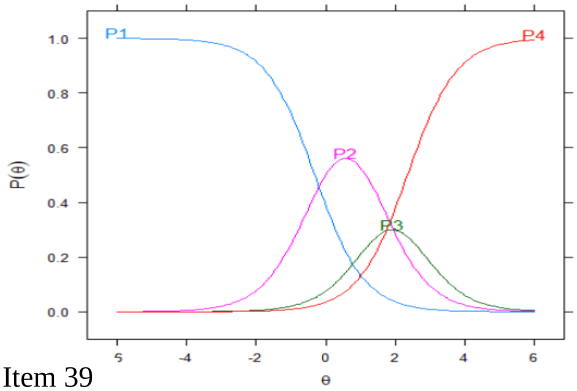
Item 36



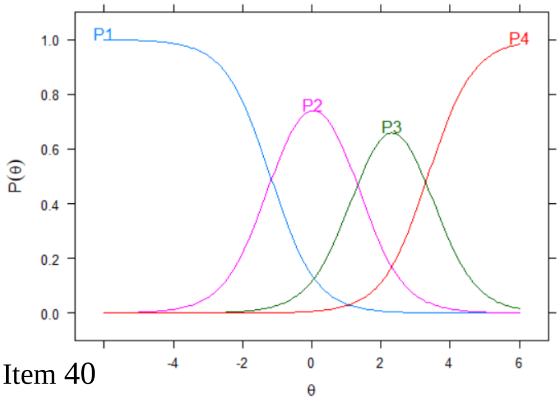
Item 37



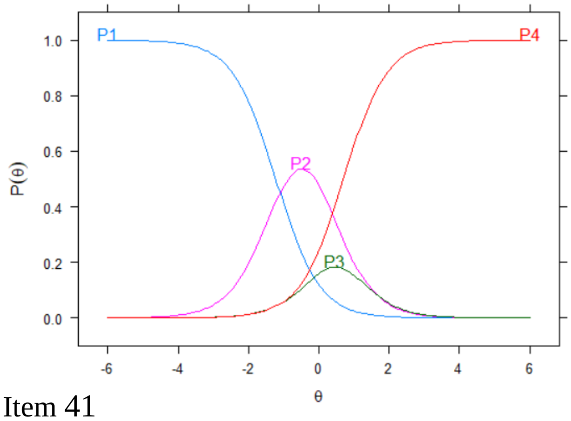
Item 38



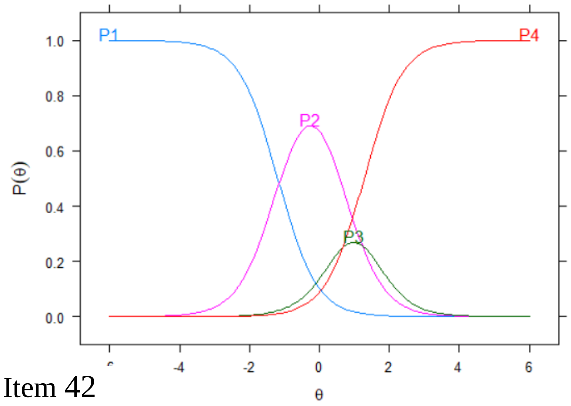
Item 39



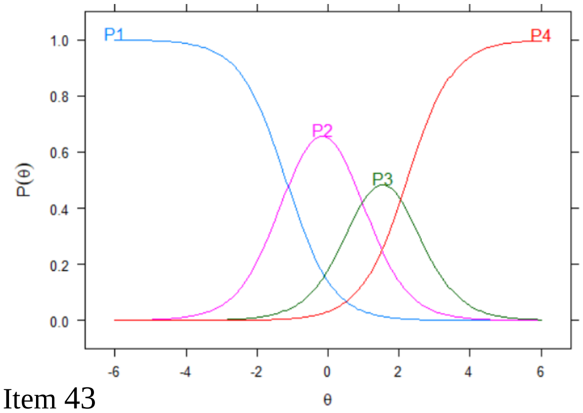
Item 40



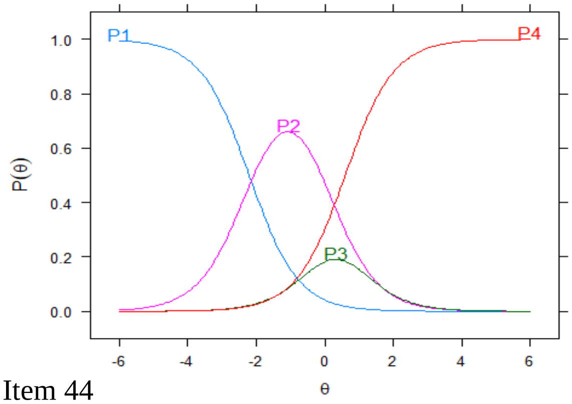
Item 41



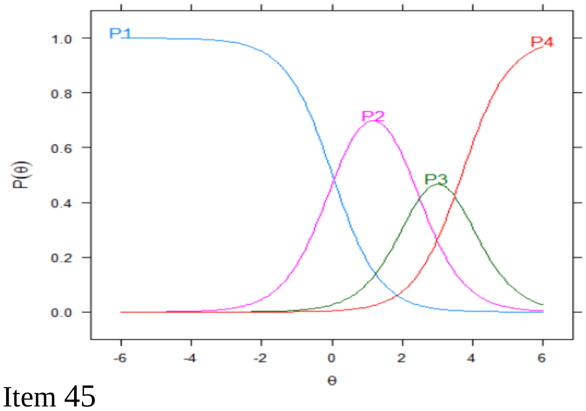
Item 42



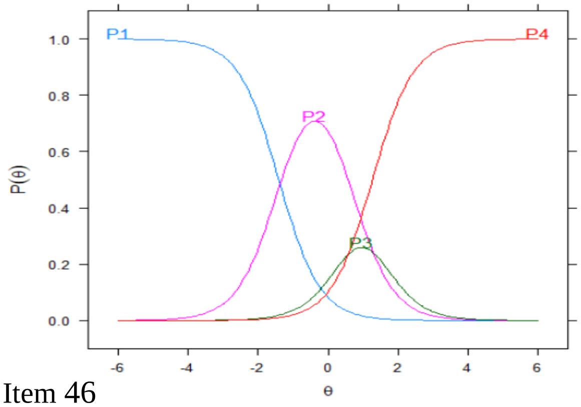
Item 43



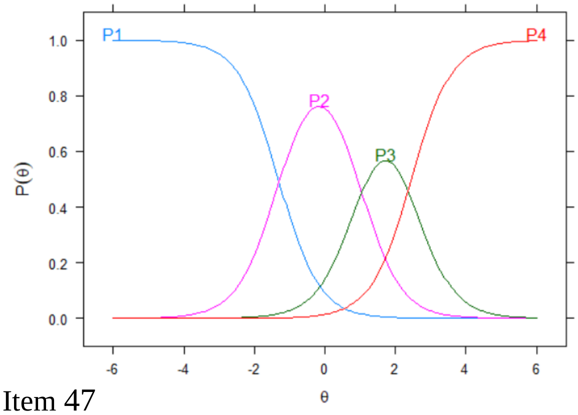
Item 44



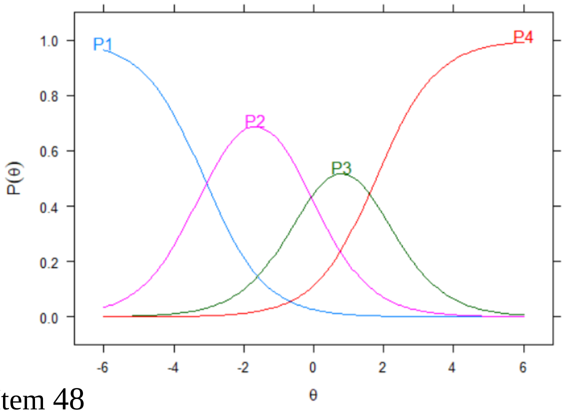
Item 45



Item 46



Item 47



Item 48

Appendix 4.11 Correlations (Significance) Among Item Parameters and Item Characteristics on the GRM Model of 4-Categroy EI Accuracy Scale

	Discrimination	Difficulty	Difficulty	Difficulty	Difficulty	Discrimination	Distance	Distance
	$(a_{\_4\text{-Category}})$	(b_1)	(b_2)	(b_3)	(b_{overall})	$(a_{\_3\text{-Category}})$	$(b_2 - b_1)$	$(b_3 – b_2)$
$a_{\_4\text{-Category}}$	1							
b_1	0.24 (0.099)	1						
b_2	-0.04 (0.770)	0.79 (0.000)	1					
b_3	-0.19 (0.185)	0.63 (0.000)	0.87 (0.000)	1				
b_{overall}	-0.03 (0.853)	0.85 (0.000)	0.97 (0.000)	0.93 (0.000)	1			
$a_{\_3\text{-Category}}$	0.94 (0.000)	0.11 (0.439)	-0.13 (0.390)	-0.24 (0.094)	-0.12 (0.431)	1		
$b_2 - b_1$	-0.42 (0.003)	-0.19 (0.197)	0.45 (0.001)	0.48 (0.001)	0.31 (0.033)	-0.37 (0.010)	1	
$b_3 – b_2$	-0.32 (0.025)	0.08 (0.595)	0.25 (0.086)	0.69 (0.000)	0.42 (0.003)	-0.30 (0.040)	0.29 (0.047)	1
P1 (%)	0.11 (0.445)	0.96 (0.000)	0.80 (0.000)	0.66 (0.000)	0.86 (0.000)	0.01 (0.949)	-0.12 (0.407)	0.13 (0.394)
P2 (%)	0.10 (0.488)	-0.28 (0.053)	0.15 (0.300)	0.20 (0.166)	0.06 (0.705)	0.09 (0.543)	0.66 (0.000)	0.18 (0.228)
P3 (%)	-0.18 (0.223)	-0.55 (0.000)	-0.55 (0.000)	-0.15 (0.294)	-0.42 (0.003)	-0.08 (0.606)	-0.07 (0.636)	0.51 (0.000)
P4 (%)	-0.11 (0.437)	-0.64 (0.000)	-0.79 (0.000)	-0.89 (0.000)	-0.86 (0.000)	-0.04 (0.798)	-0.33 (0.022)	-0.61 (0.000)
TPE (P1)	0.23 (0.114)	1.00 (0.000)	0.80 (0.000)	0.64 (0.000)	0.86 (0.000)	0.11 (0.473)	-0.18 (0.219)	0.09 (0.561)
TPE (P2)	-0.39 (0.007)	-0.22 (0.129)	0.42 (0.003)	0.47 (0.001)	0.28 (0.054)	-0.34 (0.020)	0.99 (0.000)	0.31 (0.031)
TPE (P3)	-0.27 (0.064)	-0.02 (0.881)	0.12 (0.403)	0.59 (0.000)	0.30 (0.041)	-0.23 (0.108)	0.23 (0.115)	0.98 (0.000)
TPE (P4)	0.15 (0.324)	-0.63 (0.000)	-0.86 (0.000)	-1.00 (0.000)	-0.92 (0.000)	0.20 (0.176)	-0.45 (0.001)	-0.71 (0.000)
Illustrated Practical Value (IPV, P3)	-0.11 (0.541)	0.21 (0.353)	0.24 (0.249)	0.59 (0.000)	0.39 (0.006)	-0.05 (0.708)	0.22 (0.625)	0.85 (0.000)

	Response (%) Category 1 (P1)	Response (%) Category 2 (P2)	Response (%) Category 3 (P3)	Response (%) Category 4 (P4)	TPE (P1)	TPE (P2)	TPE (P3)	TPE (P4)
P1 (%)	1							
P2 (%)	-0.39 (0.007)	1						
P3 (%)	-0.53 (0.000)	0.02 (0.911)	1					
P4 (%)	-0.62 (0.000)	-0.33 (0.023)	0.05 (0.758)	1				
TPE (P1)	0.97 (0.000)	-0.29 (0.047)	-0.55 (0.000)	-0.64 (0.000)	1			
TPE (P2)	-0.17 (0.261)	0.72 (0.000)	-0.05 (0.723)	-0.34 (0.018)	-0.21 (0.143)	1		
TPE (P3)	0.02 (0.915)	0.20 (0.165)	0.63 (0.000)	-0.56 (0.000)	-0.02 (0.917)	0.26 (0.070)	1	
TPE (P4)	-0.66 (0.000)	-0.23 (0.114)	0.12 (0.398)	0.92 (0.000)	-0.64 (0.000)	-0.45 (0.001)	-0.61 (0.000)	1
IPV (P3)	0.21 (0.241)	0.11 (0.979)	0.4 (0.005)	-0.58 (0.000)	0.21 (0.324)	0.21 (0.573)	0.86 (0.000)	-0.6 (0.000)

Appendix 5.1 Frequency of the Observed and Expected Responses of Three Misfitting Items

Category	Ɵ	Frequency (Q15)			Ɵ	Frequency (Q33)			Ɵ	Frequency (Q29)		
	mean	Obs.	Exp.	z.Res.	mean	Obs.	Exp.	z.Res.	mean	Obs.	Exp.	z.Res.
1	-1.57	14	12.34	0.47	-1.37	13	13.43	-0.12	-1.34	14	14.35	-0.09
2		6	8.17	-0.76		8	8.05	-0.02		8	7.34	0.24
3		1	0.39	0.97		0	0.44	-0.66		0	0.31	-0.56
4		0	0.10	-0.32		1	0.08	3.31		N/A	N/A	N/A
1	-0.95	4	6.83	-1.08	-0.92	9	7.95	0.37	-0.92	9	8.57	0.15
2		16	11.93	1.18		11	10.97	0.01		11	10.74	0.08
3		0	0.98	-0.99		0	0.92	-0.96		0	0.69	-0.83
4		0	0.26	-0.51		0	0.17	-0.41		N/A	N/A	N/A
1	-0.67	4	4.94	-0.42	-0.68	6	5.82	0.07	-0.69	9	6.19	1.13
2		15	13.17	0.50		13	12.49	0.14		11	12.68	-0.47
3		1	1.49	-0.40		1	1.42	-0.35		0	1.13	-1.06
4		0	0.41	-0.64		0	0.27	-0.52		N/A	N/A	N/A
1	-0.40	6	3.46	1.36	-0.49	5	4.43	0.27	-0.5	4	4.57	-0.27
2		13	13.73	-0.20		15	13.22	0.49		14	13.77	0.06
3		1	2.18	-0.80		0	1.96	-1.40		2	1.66	0.26
4		0	0.63	-0.79		0	0.38	-0.62		N/A	N/A	N/A
1	-0.13	4	2.37	1.06	-0.3	6	3.27	1.51	-0.3	1	3.24	-1.25
2		10	13.57	-0.97		11	13.49	-0.68		18	14.32	0.97
3		4	3.10	0.51		3	2.69	0.19		1	2.43	-0.92
4		2	0.96	1.06		0	0.55	-0.74		N/A	N/A	N/A
1	0.03	1	1.87	-0.64	-0.08	1	2.27	-0.84	-0.09	3	2.20	0.54
2		15	13.15	0.51		15	13.17	0.50		13	14.23	-0.33
3		4	3.75	0.13		3	3.73	-0.38		4	3.57	0.23
4		0	1.23	-1.11		1	0.83	0.19		N/A	N/A	N/A
1	0.28	1	1.30	-0.26	0.23	1	1.29	-0.26	0.23	3	1.15	1.72
2		14	12.10	0.55		10	11.63	-0.48		14	12.75	0.35
3		3	4.82	-0.83		7	5.59	0.60		3	6.10	-1.26
4		2	1.78	0.17		2	1.49	0.42		N/A	N/A	N/A
1	0.66	2	0.72	1.51	0.74	0	0.50	-0.71	0.71	0	0.41	-0.64
2		8	9.70	-0.55		10	7.58	0.88		5	8.39	-1.17
3		10	6.50	1.37		8	8.36	-0.13		15	11.19	1.14
4		0	3.08	-1.76		2	3.56	-0.83		N/A	N/A	N/A
1	1.05	0	0.39	-0.62	1.05	0	0.28	-0.53	1.05	0	0.20	-0.45
2		6	6.91	-0.35		1	5.13	-1.82		4	5.32	-0.57
3		8	7.58	0.15		9	8.92	0.03		16	14.48	0.40
4		6	5.13	0.39		10	5.67	1.82		N/A	N/A	N/A
1	1.70	0	0.14	-0.37	1.79	0	0.07	-0.27	1.8	0	0.04	-0.21
2		2	3.30	-0.71		0	1.70	-1.30		1	1.46	-0.38
3		4	7.00	-1.13		9	6.45	1.00		21	20.50	0.11
4		15	10.56	1.37		13	13.78	-0.21		N/A	N/A	N/A

Note. Obs.: observed responses; Exp.: expected responses; z.Res.: z. Residuals

Appendix 5.2 Differences in Distribution of Empirical and Model-Based Scores and Z.Residuals of Three Misfitting Items by Person Ability Percentiles on the 4- and 3-Category EI Accuracy Scales

	4-category scale						3-category scale		
Ability	Item 15			Item 33			Item 29		
Percentile	Theta	Freq. (Δ)	z.Residuals	Theta	Freq. (Δ)	z.Residuals	Theta	Freq. (Δ)	z.Residuals
10% or lower	-1.57	4.53	2.51	-1.37	1.84	4.11	-1.34	1.32	0.89
11% - 20%	-0.95	8.14	3.76	-0.92	2.17	1.75	-0.92	1.38	1.06
21% - 30%	-0.67	3.65	1.96	-0.68	1.37	1.09	-0.69	5.61	2.66
31% - 40%	-0.40	5.08	3.15	-0.49	4.69	2.78	-0.50	1.15	0.60
41% - 50%	-0.13	7.15	3.60	-0.30	6.08	3.12	-0.30	7.35	3.14
51% - 60%	0.03	4.20	2.39	-0.08	4.00	1.91	-0.09	2.46	1.10
61% - 70%	0.28	4.24	1.80	0.23	3.85	1.75	0.23	6.20	3.33*
71% - 80%	0.66	9.57	5.19*	0.74	4.85	2.54	0.71	7.61	2.95
81% - 90%	1.05	2.59	1.51	1.05	8.82	4.19*	1.05	3.04	1.42
91% - 100%	1.70	8.88	3.59	1.79	5.10	2.79	1.80	1.01	0.70
Total		58.03	29.46		42.77	26.03		37.13	17.85

Notes. Freq. (Δ): subtotal of the absolute values of differences in the frequencies between observed and empirical responses at ability levels; z.Residuals: subtotal of the absolute values of z.Residuals

Appendix 5.3 Differences in Distribution of Empirical and Model-Based Scores and Z.Residuals of Three Misfitting Items by Response Categories on the 4- and 3-Category EI Accuracy Scales

Category	4-category scale				3-category scale	
	Item 15		Item 33		Item 29	
	Freq. (Δ)	z.Residuals	Freq. (Δ)	z.Residuals	Freq. (Δ)	z.Residuals
1	12.57	7.81	7.38	4.95	9.71	6.45
2	20.03	6.27	16.57	6.33	14.17	4.63
3	13.16	7.29	9.19	5.70	13.26	6.76
4	12.28	8.10	9.64	9.06	N/A	N/A
Total	58.04	29.47	42.78	26.04	37.14	17.84

Notes. Freq. (Δ): subtotal of the absolute values of differences in the frequencies between observed and empirical responses by response category levels; z.Residuals: subtotal of the absolute values of z.Residuals

Appendix 6.1 Comparison of Observed and Model-based Scores of the Response of 21 Underfitting Examinees Either on the 3-Category or 4-Cateogory Accuracy Scale

No.	ID	Score type (scale)	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
1	F1-40	observed	1	1	1 [b]	1	2	1	2	1	1	2 [a,b]	1	3 [a,b]
		model (4-category)	1.31	1.58	2.03	1.38	1.44	1.74	1.97	1.49	1.33	1.03	1.22	1.46
		model (3-category)	1.30	1.56	1.94	1.33	1.42	1.72	1.88	1.49	1.30	1.02	1.22	1.48
2	F1-65	observed	3 (4 [b])	3 (4 [b])	3	3	3	3	3 (4)	3 (4)	2	1	3 [a,b]	1 [a,b]
		model (4-category)	2.54	2.85	3.51	2.92	3.11	2.75	3.33	3.23	2.44	1.57	1.82	2.77
		model (3-category)	2.41	2.63	2.82	2.64	2.80	2.62	2.72	2.81	2.48	1.61	1.79	2.56
3	F1-73	observed	2	2	3	3	3	1 [a,b]	3	3	3	1	3 [a,b]	1 [a,b]
		model (4-category)	2.09	2.30	3.10	2.35	2.60	2.41	2.89	2.68	2.08	1.26	1.61	2.27
		model (3-category)	2.19	2.39	2.71	2.41	2.63	2.47	2.59	2.65	2.23	1.39	1.68	2.35
4	F1-80	observed	3 [a,b]	2	1 [a,b]	3 (4 [b])	1 [a,b]	3	1 [a,b]	3	3 [b]	1	1	2
		model (4-category)	2.04	2.26	3.05	2.29	2.54	2.37	2.84	2.62	2.04	1.24	1.58	2.23
		model (3-category)	2.04	2.23	2.62	2.22	2.47	2.34	2.48	2.49	2.08	1.26	1.59	2.20
5	F1-137	observed	3 (4 [b])	3 (4 [b])	3 (4)	3	3	3	3 (4)	1 [a,b]	2	3 [a,b]	2	1 [a,b]
		model (4-category)	2.53	2.84	3.51	2.91	3.10	2.75	3.32	3.22	2.43	1.57	1.82	2.76
		model (3-category)	2.39	2.61	2.81	2.62	2.79	2.61	2.71	2.80	2.46	1.59	1.78	2.54
			Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24
6	F2-20	observed	2	3	3 (4 [b])	3	3	3	1 [a,b]	1 [a,b]	2	1 [a,b]	2	2
		model (4-category)	1.55	2.67	2.41	2.49	2.13	3.24	2.82	2.97	1.83	2.02	1.75	2.17
		model (3-category)	1.64	2.68	2.34	2.40	2.11	2.86	2.52	2.65	1.88	2.03	1.79	2.21
7	F2-21	observed	3 [a,b]	2	3	3	1	3 (4)	1 [a,b]	2	1	1	2	2
		model (4-category)	1.36	2.52	2.25	2.30	1.92	3.05	2.59	2.67	1.71	1.90	1.64	2.02
		model (3-category)	1.36	2.47	2.16	2.19	1.84	2.71	2.31	2.39	1.69	1.87	1.62	1.98
8	F2-62	observed	2	3	3	3	3	3	1 [a,b]	3 (4)	3 (4 [b])	3	1 [a,b]	3 (4 [b])
		model (4-category)	2.06	3.09	2.92	3.07	2.80	3.65	3.39	3.64	2.19	2.36	2.06	2.77
		model (3-category)	2.11	2.92	2.68	2.76	2.62	2.97	2.81	2.92	2.27	2.36	2.13	2.74
9	F2-72	observed	1	3	1 [a,b]	1 [a,b]	1	3 (4 [b])	1 [a,b]	3 (4 [b])	2	1	1	3 [a,b]
		model (4-category)	1.16	2.30	2.05	2.06	1.64	2.75	2.25	2.26	1.52	1.73	1.48	1.78
		model (3-category)	1.18	2.27	2.02	2.03	1.62	2.53	2.12	2.15	1.54	1.74	1.48	1.80
10		observed	3 [a,b]	3	2	3	2	3	3 (4 [b])	1 [a,b]	2	1a	1	2

	F2-132	model (4-category)	1.50	2.63	2.37	2.44	2.07	3.19	2.76	2.89	1.80	1.99	1.72	2.13
		model (3-category)	1.58	2.64	2.30	2.36	2.05	2.84	2.48	2.60	1.84	2.00	1.75	2.16
			Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36
11	F3-22	observed	2	1	1	1	2	1	2	1 [b]	3 (4) [a,b]	1	1	1
		model (4-category)	1.09	1.09	1.74	1.51	1.50	1.65	1.95	2.11	1.54	1.27	1.60	1.30
		model (3-category)	1.08	1.10	1.71	1.54	1.48	1.55	1.91	1.98	1.50	1.28	1.57	1.32
12	F3-63	observed	2	2	1 [a,b]	2	3 (4 [b])	3 (4 [b])	1 [a,b]	2	1 [a,b]	1	3	2
		model (4-category)	1.54	1.49	2.15	1.97	2.19	2.50	2.67	2.85	2.15	1.87	2.16	1.70
		model (3-category)	1.54	1.49	2.10	1.97	2.07	2.15	2.40	2.52	2.12	1.87	2.11	1.69
13	F3-71	observed	1	1	2	1a	1 [a,b]	3 (4 [b])	3	3 (4 [b])	2	3 [a,b]	3	1
		model (4-category)	1.59	1.53	2.17	1.99	2.23	2.55	2.71	2.90	2.18	1.89	2.19	1.72
		model (3-category)	1.61	1.55	2.14	2.02	2.12	2.21	2.44	2.57	2.18	1.91	2.16	1.73
14	F3-72	observed	1	2	2	1	3 (4 [b])	3 (4 [b])	2	3	1 [a,b]	1	3	1
		model (4-category)	1.43	1.39	2.08	1.89	2.06	2.34	2.54	2.71	2.04	1.79	2.06	1.62
		model (3-category)	1.47	1.43	2.06	1.93	2.01	2.09	2.35	2.47	2.06	1.82	2.07	1.65
15	F3-74	observed	2	2	2	2	3 (4) [a,b]	2	2	1 [a,b]	1 [b]	1	2	2
		model (4-category)	1.40	1.37	2.06	1.87	2.04	2.31	2.51	2.69	2.02	1.77	2.05	1.61
		model (3-category)	1.37	1.35	2.01	1.87	1.93	2.00	2.28	2.39	1.97	1.75	2.00	1.59
16	F3-92	observed	2	1 [a,b]	2	1 [a,b]	3	3 (4)	2 [b]	3	3	3	3 (4 [b])	3
		model (4-category)	2.18	2.04	2.53	2.31	2.92	3.19	3.25	3.43	2.70	2.16	2.81	2.02
		model (3-category)	2.25	2.11	2.48	2.37	2.61	2.63	2.74	2.85	2.68	2.22	2.72	2.05
			Q37	Q38	Q39	Q40	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48
17	F4-03	observed	2	1 [a,b]	2	3 (4 [b])	3	3	3	2 [b]	2	3 (4 [b])	2	3 (4 [b])
		model (4-category)	2.28	2.78	2.22	2.23	3.18	2.77	2.52	3.32	1.86	2.88	2.38	2.93
		model (3-category)	2.13	2.51	2.14	2.22	2.62	2.51	2.48	2.77	1.80	2.54	2.33	2.72
18	F4-63	observed	1	1	1	1	1	3 (4) [a,b]	1	1 [b]	2	1	2	3
		model (4-category)	1.46	1.56	1.26	1.45	1.66	1.48	1.45	2.04	1.14	1.61	1.48	2.09
		model (3-category)	1.45	1.55	1.23	1.44	1.62	1.44	1.42	1.93	1.17	1.63	1.45	2.05
19	F4-70	observed	3a	3 (4 [b])	3 [a,b]	2	3 (4 [b])	1 [a,b]	2	3 (4 [b])	1	2	2	1 [a,b]
		model (4-category)	2.06	2.42	1.94	2.04	2.78	2.39	2.23	3.00	1.65	2.48	2.15	2.71
		model (3-category)	1.93	2.22	1.83	2.00	2.36	2.20	2.16	2.56	1.58	2.26	2.07	2.54
20		observed	3a	3	1	1 [a,b]	3 (4 [b])	2	3	3 (4 [b])	1	1 [a,b]	3	2

	F4-118	model (4-category)	2.02	2.36	1.89	2.01	2.72	2.33	2.19	2.94	1.62	2.42	2.11	2.67
		model (3-category)	1.94	2.24	1.85	2.01	2.38	2.22	2.18	2.58	1.59	2.28	2.08	2.55
21	F4-142	observed	2	3 (4 [b])	1	1	2	1 [a,b]	1 [b]	3	2	3	2	3 (4 [b])
		model (4-category)	1.90	2.18	1.74	1.91	2.47	2.14	2.03	2.73	1.49	2.23	1.99	2.54
		model (3-category)	1.81	2.07	1.66	1.88	2.19	2.02	1.99	2.42	1.46	2.11	1.94	2.42

Notes: [a] flagged on the 3-category scale only; [b] flagged on the 4-category scale only; [a,b] flagged on both 3- and 4-category scales

Appendix 6.2 Potential Sources of Misfitting Responses

No.	ID	Item No.	Prompt (Changed words in bold)	Response (Changed words in bold)	Observed (expected) score	GR errors	Words, phrase, or items with issues	Other sources
Main source A. Paraphrasing using simpler language (less complex syntax and more frequent lexis)								
1	F1-40	Q10	Purdue **ranks** second in foreign **student enrollment** among all public schools.	Purdue **has second** in all public universities for **getting** foreign **students**.	2 (1)	minor	*has (vs. rank) getting foreign students (vs. foreign student enrollment)*	C2
Main source B. Omission of semantically less essential lexis								
2	F1-80	Q1	Most students declare their major at the end of their sophomore year **in college.**	Most students declare their major at the end of their sophomore year	3(2)	none	*in college*	
3	F1-80	Q3	Last month, we traveled to Chicago, which is the **third** largest city **in the country**.	Last month, we traveled to Chicago, which is the largest city.	1 (3)	none	*third, in the country*	C2
4	F1-73	Q11	Taking a part-time job on campus has been shown to **help** students **succeed in college.**	Taking a part-time job on campus has been shown to **be helpful for** students.	3 (2)	none	*succeed in college*	C1
5	F1-40	Q12	Although **he** did not **review** for the final exam, he scored very high **on that test**.	Although **the student** didn't **study** for the final exam, he scored **really** high.	3 (2)	none	*on that test*	C1
6	F3-71	Q34	Students can take courses that **have nothing to do with** their major **areas of study**.	Students can take courses that **are not related to** their major.	3 (2)	none	*areas of study*	A, C1
7	F4-118	Q37	As you can see on the course schedule, we will not have a final exam **for this course**.	As you can see on the course schedule, we will not have a final exam.	3 (2)	none	*for this course*	C1
Main source C1. Rating inconsistency regarding semantic judgement (paraphrase vs. minor deviation)								

8	F4-70	Q37	As you can see on the course schedule, we will not have **a** final exam for this course.	As you can see on the course schedule, we will not have **the** final exam for this course.	3 (2)	none	*the final exam* (vs. *a final exam*)	-
9	F4-70	Q39	Regular exercise is extremely **important** for long-term health and well-being.	Regular exercise is extremely **well** for long-term health and well-being.	3 (2)	none	*well* (vs. *important*)	-
10	F4-03	Q44	It’s hard to express your ideas if your language **skills are** low.	It's hard to express your ideas if your language **level is** low.	2 (3*)	none	*language level* (vs. *language skills*)	
Main source C2. Rating inconsistency regarding semantic judgement (minor vs. major deviation)								
11	F1-73	Q6	It doesn’t matter if you work alone or in a group **on your homework**.	It doesn’t matter if you work in a group or alone **for this project**.	1 (3)	none	*for this project* (vs. *on your homework*)	-
12	F2-72	Q15	Joining a **student** club on campus is a great way to improve your social **skills**.	Joining a club on campus is a great way to improve your social.	1 (2)	none	*club* (vs. *student club*), *social* (vs. *social skills*)	-
13	F2-72	Q16	The wonderful thing about **English teachers** is that they know **their** students quite well.	The wonderful thing about **the teacher** is that they know students quite well.	1 (2)	minor	*the teacher* (vs. *English teachers*), *students* (vs. *their students*)	-
14	F2-20	Q19	Working part-time will help you **develop** time **management skills**.	Working part-time will help you **manage time development**.	1 (3)	none	*manage time development* (vs. *develop time management skills*)	-
15	F2-62	Q19	Working part-time will help you develop **time management** skills.	Working part-time will help you develop your **part-time managing** skills.	1 (3)	none	*part-time managing skills* (vs. *time management skills*)	D
16	F2-132	Q20	If you have a morning class, you **should** go to bed early **the night before**.	If you have a morning class, you should **have to** go to bed early **before the day of class**.	1 (3)	minor	before the day of class (vs. *the night before*)	-
17	F2-20	Q22	You can tell me what questions you have on the final project **during my office hours**.	You can tell me what questions you have on the final project **on the day of the final exams**.	1 (2)	none	*on the day of the final exams* (vs. *during my office hours*)	-

18	F3-92	Q26	**Foreign** students are only permitted to work part-time **for employers** on campus.	Students are only permitted to work **on** part-time jobs on campus	1 (2)	minor	*Students* (vs. *foreign students*)	-
19	F3-71**	Q28	When you look at the **course** schedule, you **will** see the **dates for** midterm and final **exams**.	When you look at the schedule, you **can** see the midterm and final.	1 (2**)	none	*course, dates*	-
20	F3-92	Q28	When you look at the **course** schedule, you **will see** the **dates for** midterm and final exams.	When you look at the schedule, you **can look for** the midterms and final exams.	1 (2)	none	*course, dates*	-
21	F3-63	Q33	The senior student was **talking** about **his own** story **of** finding an apartment.	The senior student was **telling a** story **about** finding an apartment.	1 (2)	none	*his own* (vs. *a*)	-
22	F4-70	Q42	**Students** who enjoy **working** in groups are **more** likely to succeed.	**Student** enjoying **work** in groups are likely to succeed.	1 (2)	minor	*likely* (vs. *more likely*)	-
23	F4-142	Q42	Students who **enjoy** working in groups are more likely to succeed.	Students who work in groups are more likely to succeed.	1 (2)	none	*working* (vs. *enjoy working*)	-
24	F4-63*	Q44	It's hard to express your ideas if your **language skills** are low.	It is hard to express your ideas if your **expression level** is low.	1 (2*)	none	*expression level* (vs. *language skills*)	-
25	F4-118	Q46	**Meeting people and** making friends should be an **important** part of your college life.	Making **people as your** friends should be a part of your college life.	1 (2)	minor	*meeting people*	-
Main source D. Major semantic deviation with no/few errors and high similarity								
26	F1-80	Q5	First of all, you must attend all the classes to pass this **course**.	First of all, you must attend all the classes to pass this **score**.	1 (3)	none	*score*	-
27	F1-137	Q8	If you record your lectures, you can **revise** your class notes later.	If you record your lectures, you can **review** your class notes later.	1 (3)	none	*review*	-
28, 29	F1-73	Q12	Although he did not **review** for the final exam, he scored very high **on that test**.	Although he did not **study** for the final exam, he scored very high **for this class**.	1 (2)	none	*for this class* (vs. *on that test*)	C2

30	F1-137	Q12	Although he did not review for the final exam, he scored very high on that **test**.	Although he did not review for the final exam, he scored very high on that **class**.	1 (3)	none	*class* (vs. *test*)	C2
31	F2-21, F2-72	Q19	Working part-time will help you develop **time** management skills.	Working part-time will help you develop management skills.	1 (2)	none	*management* (vs. *time management*)	C2
32	F2-20	Q20	If you have a morning class, you should go to bed early **the night before**.	If you have a morning class, make sure you go to bed early **that day**.	1 (3)	none	*that day* (vs. *the night before*)	-
33	F2-62	Q23	Borrowed books from the library must be **returned** or renewed **by** the **posted** due **dates**.	Borrowed books from the library must be renewed or **posted from** the given due **date**.	1 (2)	minor	*posted from* (vs. *renewed by*)	C2
34	F3-63	Q31	You can also talk to senior **students** about selecting courses.	You can also talk to senior **people** about selecting courses.	1 (2)	none	*people* (vs. *students*)	-
35	F3-22	Q32	Many students live **off** campus because the rent is much lower.	Many students live **on** campus because rent is much lower.	1 (2)	none	*on* (vs. *off*)	-
36	F3-74	Q33	The senior student **was** talking about his **own story of** finding an apartment.	The senior student **is** talking about his **experiment about** finding an apartment.	1 (2)	minor	*his experiment (vs. his own story)*	-
37	F4-03	Q38	In the event of a **car** accident, you should first stay calm and then call the police.	In the event of a **fire** accident, you should first stay calm and then call the police.	1 (3)	none	*car accident* (vs. *fire accident*)	-
38	F4-142*	Q43	Most college students move out of **the** dorms after their **sophomore** year.	Most college students move out of **their** dorms after their **first** year.	1 (2*)	none	*First year* (vs. *sophomore year*)	-
39	F4-70	Q48	You should talk to your advisor **if** you are not sure **what** courses to take next semester.	You should talk to your advisor **that** you are not sure **which** courses to take next semester.	1 (3)	none	*that* (vs. *if*)	C2

Main source E. Rating inconsistency regarding the degree of grammatical error: minor vs. major

40	F3-74	Q32	Many **students** live **off** campus because the rent is much lower.	Many **student** live **out of** campus because the rent is much **more** lower.	1 (2)	minor	*much more lower* (vs. *much lower*)	-
41	F3-72	Q33	The senior **student was** talking about **his own** story **of finding an** apartment.	The senior **students is** talking about story **that** he **choose his own** apartment.	1 (2)	minor	*is* (vs. *was*)	-
<u>**Main source F**. No rating/scale issue: impacts of other misfitting/discrepant responses within the examinee</u>								
42	F1-40	Q3	Last month, we **traveled to** Chicago, **which** is the **third** largest city **in the country**.	Last month, we **visited to** Chicago, **where** is the largest city.	1 (2)	minor	Q10, Q12 (overestimated)	-
43	F1-80	Q9	The way that English classes are taught here might differ from the way in your country.	The way that English classes are taught here might differ from the ways they are taught in your country.	3 (2)	none	Q3, Q5, (underestimated)	
44	F1-137	Q10	Purdue ranks second in foreign student enrollment **among** all public schools.	Purdue ranks second in foreign student enrollment **out of** all public schools.	3 (2)	none	Q8, Q12, (underestimated)	-
45	F2-21	Q13	The amount of work involved in studying for final exams can **overwhelm you**.	The amount of work involved in studying for final exams can **be overwhelming**.	3 (1)	none	Q19 (underestimated)	-
46	F2-132	Q13	The amount of work involved in studying for **final exams can** overwhelm you.	The amount of work involved in studying for **finals may** overwhelm you.	3 (2)	none	Q20 (underestimated)	-
47	F2-72	Q24	After he worked on the project all evening, the student went **directly** to bed.	After he worked on the project all evening, the student went **right** to bed.	3 (2)	none	Q15, Q16, Q19 (underestimated)	-
48	F3-71	Q29	**Regular workouts benefit the body as well as the mind.**	**Working out benefits your brain.**	1 (2)	minor	Q34 (overestimated)	-

49	F4-118	Q40	In other words, you must submit all your homework assignments **on the course website.**	In other words, you must submit all **coursework** and homework.	1 (2)	none	Q37 (overestimated)	-
Source G. No rating/scale issue: unknown examinee-related sources								
50	F1-80	Q7	**Earning** money is the main reason for students to get a job.	**Paying in** money is the main reason for students to **go** to a **university**.	1 (2)	none	-	-
51	F1-65	Q11	Taking a part-time job on campus has been shown to help students **succeed** in college.	Taking a part-time job on campus has been shown to help students **to succeed** in college.	3 (2)	none	-	-
52	F1-65	Q12	Although **he** did not **review for** the **final** exam, **he scored very high on that test**.	Although **he** didn't **entirely review** the test.	1 (3)	major	-	-
53	F2-132**	Q22	You can tell me what **questions** you have **on the final project** during my office hours.	You can tell me what **project do** you have during my office hours.	1 (2**)	minor	-	-
54	F3-63	Q27	Students can keep **the** books **that they borrow from the library for a** semester.	Students can keep **their** books **until the end of the** semester.	1 (2)	minor	-	-
55	F3-92	Q31	You can also talk to senior students **about selecting** courses.	You can also talk to senior students to **ask about the** courses.	2 (3*)	none	-	-

Notes: GR=grammar; *flagged on 4-category scale only; **flagged on 3-category scale only

Appendix 7.1 Correlations Between Item Discrimination Across the Eight Scales

	Binary	3C	3C-FSD	4C-FGD	4C-FSGD	4C	4C-FSD	5C-FGD	5C-FSGD
Binary	1								
3C	0.67	1							
3C-FSD	0.57	0.82	1						
4C-FGD	0.59	0.96	0.79	1					
4C-FSGD	0.52	0.72	0.89	0.79	1				
4C	0.57	0.9	0.74	0.89	0.66	1			
4C-FSD	0.46	0.7	0.9	0.71	0.83	0.8	1		
5C-FGD	0.52	0.88	0.71	0.93	0.73	0.96	0.79	1	
5C-FSGD	0.43	0.6	0.77	0.69	0.91	0.7	0.89	0.79	1

Appendix 7.2 Correlations Between Item Difficulty (Threshold of the Paraphrase Category) Across the Eight Scales

	Binary	3C	3C-FSD	4C-FGD	4C-FSGD	4C	4C-FSD	5C-FGD	5C-FSGD
Binary	1								
3C	0.95	1							
3C-FSD	0.93	0.98	1						
4C-FGD	0.95	0.99	0.99	1					
4C-FSGD	0.94	0.98	1	0.99	1				
4C	0.93	0.99	0.98	0.99	0.98	1			
4C-FSD	0.93	0.98	1	0.99	0.99	0.98	1		
5C-FGD	0.95	0.99	0.99	1	0.99	0.99	0.99	1	
5C-FSGD	0.94	0.98	1	0.99	1	0.98	1	1	1

Appendix 9.1 Ranger best fitting model (mtree = 300)

Rank	Mtry	Node size	Sampe size	OOB error (RMSE)
Outcome : Difficulty at all levels				
1	35	3	0.8	0.211343
2	31	3	0.8	0.212777
3	33	3	0.8	0.212799
4	29	3	0.8	0.213544
5	25	3	0.8	0.216243
6	27	3	0.8	0.21637
7	33	3	0.7	0.219793
8	35	5	0.8	0.220355
9	31	5	0.8	0.220518
10	35	3	0.7	0.221589
Outcome : Difficulty at the paraphrase level				
1	31	3	0.8	0.1581
2	33	3	0.8	0.15853
3	25	3	0.8	0.15857
4	33	5	0.8	0.15873
5	35	3	0.8	0.15874
6	35	5	0.8	0.15886
7	31	5	0.8	0.15902
8	29	3	0.7	0.15925
9	27	3	0.7	0.15937
10	29	3	0.8	0.15942
Outcome : Difficulty at the minor error/deviation level				
1	35	5	0.8	0.306057
2	35	3	0.8	0.306232
3	33	3	0.8	0.306975
4	35	3	0.7	0.306981
5	31	5	0.632	0.307011
6	35	5	0.7	0.307207
7	35	3	0.632	0.307611
8	35	5	0.632	0.307715
9	35	7	0.8	0.307877
10	31	5	0.8	0.308366

Appendix 9.2 Comparison of RF Model Performance

Model	Model specification				Performance (training set)				Performance (testing set)		
	Mtree	Mtry	node size	sample fraction	OBB MSE	OOB RMSE	Var explained	cor pred-vs data	MSE	RMSE	cor pred-vs data
Outcome: item difficulty at all levels											
RFb-B	300	17	5	63.2%	0.0549	0.2344	96.96%	0.9881	0.0450	0.2120	0.9782
RFb	300	24	5	63.2%	0.0395	0.1987	97.81%	0.9923	0.0333	0.1824	0.9824
RGb	300	35	3	80.0%	0.0447	0.2113	97.53%	0.9920	0.0400	0.2001	0.9778
SRCb	300	24	5	63.2%	0.0520	0.2280	97.20%	0.9854	0.0462	0.2150	0.9751
diff para											
RFbP-B	300	16	5	63.2%	0.0280	0.1673	95.02%	0.9774	0.0247	0.1573	0.9622
RFbP	300	11	5	63.2%	0.0265	0.1627	95.29%	0.9747	0.0245	0.1566	0.9616
RGbP	300	31	3	80.0%	0.0235	0.1533	95.83%	0.9848	0.0212	0.1458	0.9654
SRCbP	300	11	5	63.2%	0.0272	0.1649	95.47%	0.9677	0.0233	0.1527	0.9569
diff mnr											
RFbM-B	300	16	5	63.2%	0.0469	0.2165	90.85%	0.9708	0.0501	0.2239	0.9316
RGbM	300	35	5	80.0%	0.0934	0.3056	83.40%	0.8960	0.0649	0.2548	0.8940
SRCbM	300	16	5	63.2%	0.0832	0.2885	85.32%	0.9172	0.0875	0.2958	0.8409

Appendix 9.3 Variable Importance: Increased MSE and Node Purity When a Variable Replaced by Random Noise (Variable Presented in the Alphabetical Order)

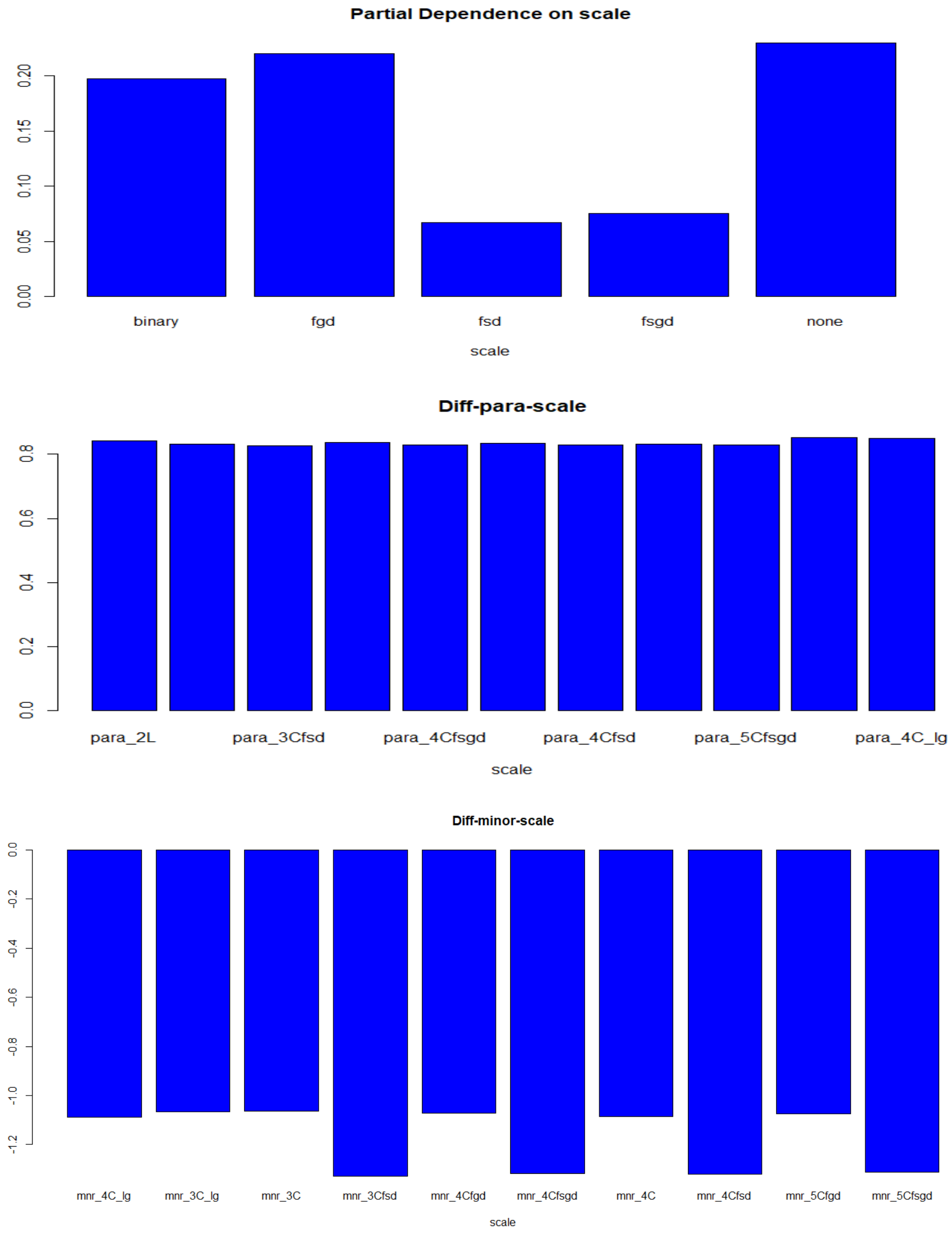
Ling Features	All				Paraphrase				Minor Error/Deviation			
	Increased MSE			Node Purity	Increased MSE			Node Purity	Increased MSE			Node Purity
	Z	abs.	Rank		Z	abs.	Rank		Z	abs.	Rank	
Article A	7.22	0.01	31	1.87	4.63	0.00	34	0.43	3.77	0.00	36	0.30
Article The	9.12	0.01	25	4.17	6.61	0.01	23	1.16	8.28	0.01	19	1.31
Association strength, COCA magazine, 3-1-to-2gram (DP)	23.00	0.06	3	23.98	13.79	0.07	1	7.71	15.71	0.09	2	14.23
Association strength, COCA spk. 3-2-to-1gram (DP)	19.85	0.07	4	25.94	11.78	0.05	4	7.15	12.56	0.06	6	7.21
av_nsubj_deps_NN	4.92	0.01	42	3.17	5.35	0.01	32	1.25	5.10	0.00	29	0.43
Average faith score construction (cue), verb, COCA fiction	16.50	0.05	7	18.16	10.96	0.03	7	4.11	11.20	0.04	10	3.67
AWL_Sublist_8_Normed	6.04	0.00	36	2.87	3.83	0.00	38	0.43	5.72	0.00	25	0.69
CN_C	8.08	0.01	27	4.34	6.23	0.03	24	3.61	5.11	0.01	28	1.27
comp	2.55	0.00	47	0.43	4.50	0.00	35	0.23	2.50	0.00	43	0.08
Deps. per nominal (std.)	11.24	0.05	19	18.65	8.39	0.06	17	6.62	8.38	0.01	18	1.87
Deps. per obj. of prep. (no pronouns, std.)	12.21	0.15	17	37.46	10.97	0.13	6	14.29	9.32	0.10	14	8.42
Deps. per obj. of prep. (std.)	11.11	0.13	20	33.15	11.15	0.14	5	13.84	10.13	0.12	11	8.96
Determiners	15.90	0.05	9	15.57	10.11	0.05	10	4.81	11.26	0.06	9	5.62
ingGrnd	4.45	0.01	45	2.07	4.14	0.01	37	0.97	2.69	0.00	42	0.36
ingProg	5.50	0.00	39	0.36	2.50	0.00	44	0.19	3.44	0.00	39	0.11
Lexical density (tokens)	15.71	0.04	11	13.88	8.05	0.05	18	5.72	9.89	0.03	13	3.70
mtld_original_aw	10.09	0.01	22	4.67	6.61	0.01	22	1.30	6.55	0.01	23	1.06
nouns_as_modifiers	9.41	0.03	24	10.35	5.97	0.02	26	2.26	8.40	0.02	17	2.48

NP elaboration	17.18	0.26	6	84.39	12.07	0.18	3	21.18	13.55	0.14	4	13.17
Num. of syllables	15.24	0.15	13	50.48	9.34	0.10	14	10.15	6.35	0.03	24	3.35
Past	7.32	0.00	30	2.01	5.44	0.00	31	0.81	2.32	0.00	45	0.08
Past participle / perfect	14.31	0.06	14	28.70	7.45	0.07	20	9.49	13.33	0.07	5	12.37
Plurals	10.77	0.02	21	7.43	10.00	0.04	12	3.91	4.49	0.01	31	0.97
possessives	11.46	0.01	18	5.70	7.09	0.01	21	0.90	6.81	0.01	22	1.44
Pperfect	5.37	0.01	41	7.64	8.45	0.05	16	10.92	5.41	0.01	26	1.65
Prep. per obj. of prep. (no pronouns)	7.12	0.04	32	13.33	7.73	0.06	19	5.48	10.00	0.11	12	10.97
prep_about	4.48	0.00	44	0.59	4.67	0.00	33	0.50	1.31	0.00	48	0.04
prep_among	5.57	0.00	38	2.33	3.80	0.00	39	0.69	4.25	0.01	32	1.74
prep_at	4.81	0.00	43	2.13	3.44	0.00	41	0.47	3.07	0.00	41	0.40
prep_by	3.63	0.00	46	0.35	2.00	0.00	45	0.10	4.12	0.00	33	0.20
prep_during	2.16	0.00	48	0.04	0.00	0.00	47	0.00	2.41	0.00	44	0.02
prep_for	7.74	0.01	29	1.86	5.93	0.01	28	0.82	4.82	0.00	30	0.40
prep_in	6.21	0.01	35	2.78	5.50	0.01	30	1.11	3.96	0.01	34	1.05
prep_of	6.43	0.00	34	1.00	4.25	0.00	36	0.23	3.11	0.00	40	0.21
prep_on	7.02	0.00	33	0.87	2.60	0.00	43	0.32	3.54	0.00	38	0.18
prep_to	5.39	0.00	40	1.41	3.40	0.00	42	0.60	3.83	0.00	35	0.21
prepAll	7.86	0.01	28	2.39	3.53	0.01	40	0.76	3.72	0.01	37	1.01
rcmod_dobj_deps_struct	2.12	0.00	49	0.03	1.00	0.00	46	0.02	1.76	0.00	47	0.05
Syntactic diversity & frequency	12.91	0.08	16	22.24	9.60	0.08	13	7.87	8.91	0.02	16	2.05
Syntactic frequency	15.49	0.04	12	14.54	8.54	0.04	15	4.34	11.65	0.03	8	4.15
Third-person singular	8.68	0.02	26	8.30	5.96	0.03	27	3.59	5.19	0.01	27	0.83
Third-person singularBe	5.98	0.01	37	4.52	5.64	0.01	29	1.11	2.24	0.00	46	0.35
Type token ratio (root)	15.83	0.17	10	60.95	12.19	0.13	2	16.36	12.00	0.08	7	7.61
VAC frequency	18.98	0.05	5	16.37	10.94	0.04	8	4.43	13.84	0.06	3	6.19
Verb VAC frequency	16.33	0.11	8	33.56	10.20	0.08	9	8.27	9.19	0.05	15	5.38
VP_T	9.61	0.01	23	5.71	6.23	0.02	25	3.01	7.14	0.01	21	1.18
Word length	14.13	0.02	15	8.87	10.05	0.02	11	3.33	7.82	0.03	20	3.35

Scales / scoring methods	32.39	0.03	2	29.75	-2.67	0.00	48	1.84	26.19	0.06	1	14.27
Threshold Levels	231.82	2.43	1	1075.63								

Notes: abs.: absolute value

Appendix 9.4 Marginal Effects of Scales / Scoring Methods



Appendix 9.5 Variable importance (Outcome: Item discrimination)

	Variables (Predictors)	Increase MSE (Z score)	Increased MSE	Increased MSE(SD)	Increased Node Purity
1	Scales/scoring methods	26.540	0.046	0.002	12.277
2	Assc. strength, COCA spk. 3-2-to-1gram (DP)	20.602	0.099	0.005	8.532
3	(N) Avr. faith score construction (verb-cue, COCA fiction)	15.723	0.039	0.002	4.326
4	Assc. strength, COCA magazine, 3-1-to-2gram (DP)	13.013	0.036	0.003	3.535
5	Syntactic frequency	12.746	0.048	0.004	3.639
6	Syntactic diversity & frequency*	12.671	0.028	0.002	3.835
7	Dependents per nominal (std.)*	12.404	0.026	0.002	2.855
8	Verb VAC frequency	10.939	0.022	0.002	2.471
9	Nouns as modifiers & modifier variation*	10.830	0.015	0.001	1.900
10	VAC frequency and direct objects	10.681	0.017	0.002	2.113
11	Type token ratio (root)	9.441	0.017	0.002	1.870
12	NP elaboration	9.356	0.019	0.002	2.273
13	CN_C	9.344	0.021	0.002	1.762
14	Word length	9.104	0.011	0.001	1.282
15	Determiners	9.097	0.015	0.002	1.671
16	prep_pobj_deps_NN_struct	8.527	0.016	0.002	1.520
17	lexical_density_tokens	7.412	0.011	0.001	1.177
18	Plurals	7.288	0.005	0.001	0.808
19	pobj_NN_stdev	6.882	0.006	0.001	0.714
20	Dependents per prepositional objects (std.)	6.493	0.006	0.001	0.653
21	mtld_original_aw	6.421	0.006	0.001	0.860
22	NumSyl	6.026	0.007	0.001	0.977
23	prepAll	5.892	0.005	0.001	0.533
24	av_nsubj_deps_NN	5.744	0.009	0.002	0.985
25	VP_T	5.681	0.002	0.000	0.252
26	prep_at	5.505	0.003	0.001	0.418
27	ArticleThe	5.446	0.002	0.000	0.361
28	possessives	5.389	0.005	0.001	0.583
29	thirdSing	4.681	0.003	0.001	0.417
30	prep_for	4.661	0.004	0.001	0.502

31	ingProg	4.638	0.003	0.001	0.372
32	prep_of	4.448	0.001	0.000	0.194
33	prep_on	4.431	0.004	0.001	0.366
34	ArticleA	4.053	0.002	0.001	0.261
35	prep_to	4.011	0.001	0.000	0.198
36	PPnPerfect	3.861	0.002	0.000	0.486
37	prep_in	3.538	0.000	0.000	0.075
38	thirdSingBe	3.535	0.000	0.000	0.058
39	prep_about	3.496	0.001	0.000	0.079
40	Past	3.467	0.001	0.000	0.157
41	ingGrnd	3.378	0.001	0.000	0.122
42	prep_by	3.178	0.000	0.000	0.076
43	rcmod_dobj_deps_struct	2.713	0.000	0.000	0.132
44	AWL_Sublist_8_Normed	2.595	0.001	0.000	0.187
45	comp	2.528	0.000	0.000	0.033
46	prep_during	2.400	0.000	0.000	0.024
47	Pperfect	1.372	0.000	0.000	0.071
48	prep_among	0.783	0.000	0.000	0.019