6-25-2018

English as a Second Language Proficiency: An IRT Approach to Scoring

Merve Sarac
mervesarac1@gmail.com

English as a Second Language Proficiency: An IRT Approach to Scoring

Merve Sarac

B.S., Bogazici University, 2014
M.S., Bogazici University, 2016

A Thesis

Submitted in partial fulfillment of the

Requirements for the degree of

Master of Arts at the

University of Connecticut

2018

APPROVAL PAGE

Master of Arts Thesis

English as a Second Language Proficiency: An IRT Approach to Scoring

Presented by

Merve Sarac, B.S., M.S.,

Major Advisor _____

Eric O. Loken

Associate Advisor _____

H. Jane Rogers

Associate Advisor _____

Hariharan Swaminathan

University of Connecticut

2018

ACKNOWLEDGMENTS

There are many people who supported me through this project and to each I want to extend my sincerest gratitude. I would like to thank my adviser Dr. Eric Loken for his encouragement, enduring support, patience and immense knowledge. His guidance helped me in extending my knowledge in all the time of research. He consistently allowed this thesis to be my own work but steered me in the right the direction whenever he thought I needed it. I really appreciate his influence on my intellectual development. I also would like to thank my thesis committee members, Dr. Jane Rogers and Dr. Hariharan Swaminathan for reading my work and providing feedback with insightful comments and invaluable suggestions. I would also appreciate Merve and Muhammed's enormous help at the beginning of my research.

My sincere thanks goes to Dr. JoLynn Powers for her endless encouragement and support during my masters study. Without her support, I would not be able to recover and reach this point in my non-academic and academic life. I would like to express my special thanks to Dakota. I appreciate his endless support, guidance, and patience during my research. I would also like to thank my fiends Xiaowen, Luh, Sabine, Roman, Gozde, Mervenur, Zeynep, and my dear cousin Sedef for their friendship, sincerity, and continuous encouragement. I also thank Merve and Serpil for their precious friendship and emotional support and watching me from a distance while I worked towards my degree.

My deepest gratitude goes to my family for their unflagging love and unconditional support throughout my life and my studies. I take this opportunity to express the profound gratitude from my deep heart to my beloved father Yener, my beloved mother Zeynep, and my lovely sisters, Sule and Melike for their unconditional love, continuous support, and endless patience for the days I had to stay away from them during my research.

Lastly, I am proud of dedicating this thesis to the bright memory of my well-beloved father, Yener Sarac...

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

CLA Communicative Language Ability

CTT Classical Test Theory

DIC Deviance Information Criterion

ELPT English Language Proficiency Test

GT Generalizability Theory

ICC Item Characteristic Curve

IELTS International English Language Testing Service

IIC Item Information Curve

IRT Item Response Theory

LR Likelihood Ratio

TEPS Test of English Proficiency

TOEFL Test of English as a Foreign Language

CHAPTER 1

Introduction

English as a second language proficiency is often measured for educational and employment purposes (Graham, 1987) to determine eligibility for special programs or to make decisions regarding promotion or graduation (Abedi, 2004). Measurement of language ability with high precision and meaningful interpretations of resulting scores requires an operational definition of second language proficiency. Given the operational definition of English language proficiency (ELP) used in its measurement, test scores support the validity of claims that the scores correspond to differing levels of language proficiency.

This paper first discusses the operational definition of English proficiency as a second language through several theoretical frameworks. Next, it reviews various English language proficiency tests and statistical methods in their measurement approach. Then, it compares two different measurement approaches to scoring a summative English as a second language proficiency test. Finally, it addresses some complications of the current scoring method given the suggested measurement approach and discusses the advantages of suggested approach instead of the current approach.

The first section in this chapter gives background on the second language proficiency construct. The next section addresses several measurement approaches to second language proficiency in the context of large-scale language testing and provides examples of such national language tests. The chapter then focuses on the specific national large-scale English language proficiency test that will be analyzed.

1.1 English Language Proficiency Definition

Theoretical framework, domain specification, and empirical measurement characteristics of language tests are widely addressed issues in research on second language proficiency (Abedi, 2004, 2007; Bachman & Clark, 1987; Graham, 1987). In this regard,

various theoretical frameworks and models proposed in the language assessment literature describe language ability (Bachman, 1990; Canale & Swain, 1980; Carroll, 1968; Lado, 1961).

The first framework explaining the language proficiency is the skill/component framework (Lado, 1961), which treats language skills - speaking, listening, reading, and writing- separate from language knowledge components grammar, vocabulary, phonology. The tests developed based on this framework were linguistic in nature and referred to as discrete-point tests (Farhady, 1980). Thus, this framework contributed to matching language tests with theoretical conceptualization of language ability. However, the framework did not detail the relationship between skills and knowledge, considering whether skills are externalizations of knowledge components or if they are qualitatively distinct and did not cover the discoursal and situational contexts.

The second framework was suggested by (Carroll, 1961) and followed up by Oller (Oller, 1983). Oller proposed a psycholinguistic model of language ability in which language is regarded as integrative and unitary in nature. Thus, the development and use of integrative tests, instead of discrete-point tests, such as close and dictation became the manifestations of this holistic perspective. However, as opposed to the holistic unitary hypothesis of language ability, several researchers proposed alternative formulation of the concept of language ability. These theoretical frameworks were primarily language-oriented and did not consider its testing.

Language functions (Halliday, 1973) and the connection between text and context (Van Dijk, 1977) extended the language proficiency framework with a recognition of the context and the discourse. In modeling language ability, the communicative competence framework suggested by Canale and Swain (1980) had three main components: grammatical, socio-linguistic, and strategic competence, where competence refers to the knowledge or ability. Thus, this model extended the concept of language ability beyond linguistic competence and involved social and discourse competences. Competence in this framework is

different than performance where performance refers to the actual use of language (Hymes, 1972).

Bachman (1990) and Bachman and Palmer (1996) proposed a more comprehensive model of language ability for better testing and measurement in the context of large-scale language testing. Bachman's (1990) emphasized the need for defining language ability in a way that language test performance reflects language performance in non-test situations (Bachman, 1990). Confirming the tasks required on language tests that correspond to real-life language use, Bachman (1996) proposed a communicative language ability framework as not a complete theory of language abilities but extending the earlier models of language ability.

The theoretical framework of communicative language ability (CLA) (Bachman, 1990) consists of "..knowledge, or competence, and capacity for implementing, or executing that competence in appropriate, contextualized communicative language use." (Bachman, 1990, p. 84) In other words, ability refers both to knowledge or competence and to skill in implementing that knowledge as opposed to Farhady's (1980) and Canale and Swain's (1980) emphasis on the competence-performance distinction.

The CLA framework has three main competencies: language competence, strategic competence, and psychophysiological mechanisms (Bachman, 1990). Language competence covers specific knowledge employed in communication with language. Strategic competence is "the mental capacity for implementing the components of language competence in contextualized communicative language use." (p. 84) Psychophysiological mechanisms correspond to psychological and neurological processes employed to physically execute language, which mostly relate to the communicative performance (Canale & Swain, 1980).

This framework describes language competence as involving organizational competence and pragmatic competence Canale (1983). Organizational competence has grammatical and textual aspects involving syntax, vocabulary, cohesion, and organization. Gram-

matical aspects involve syntax governing the choice of words, their order, and their arrangement; and textual aspects involve knowledge of cohesion conventions. Pragmatic competence abilities "..include the knowledge of language functions, of sociolinguistic rules of appropriateness, and of cultural references and figurative language." (Bachman, 1990, p. 98)

The CLA framework extended the earlier definitions of strategic competence by Canale and Swain (1980) since their definitions did not describe the mechanisms which strategic competence operates. Strategic competence in Canale and Swain's (1980) framework is verbal or non-verbal communication strategies related more to either grammatical or sociolinguistic competence. However, the CLA describes strategic competence as the capacity of relating knowledge of language to knowledge structures and contextual features where communication takes place since it emphasizes dynamic interchange between context and discourse. Strategic competence matches new information to available relevant information and maps this onto efficient use of existing language abilities. For example, for inference questions in testing reading comprehension examinees must recognize what information out of its discourse is relevant to answering, and search for the relevant information in their memory.

Psychophysiological mechanisms correspond to psychological and neurological processes employed to physically execute language such as speaking, listening, reading, and writing. Reading and listening involves the receptive language use as speaking and writing involves productive language use through auditory and visual channels.

The communicative language ability framework extends the earlier models, describes the abilities of interest in language tests, and provides a means for characterizing the constructs intended to be measured. Additionally, language testing researchers have been interested in the question of whether language proficiency is a unitary competence or it is composed of distinct component traits (Bachman & Palmer, 1981; Carroll, 1983; Oller, 1976), which also has implications for language teaching and testing.

Oller (1983) who had earlier claimed that language proficiency is a general single factor, concluded that it consists of several distinct but related constructs in addition to the general language proficiency based on several studies disconfirming the unitary factor structure and highlighting its components (Bachman & Palmer, 1982; Carroll, 1983; Farhady, 1980). For example, Farhady (1980) used the Fall 1979 UCLA English as a Second Language Placement Exam (ESLPE), testing students whose native language is not English to determine whether to provide them with remedial English as a second language courses. The ESLPE has five subtests: listening and reading comprehension, grammar, cloze and dictation. Farhady (1980) found a multi-factor solution using rotation in exploratory factor analysis. Based on such research studies and the theoretical models, Bachman (1996) claimed that language proficiency is not a single trait but consists of several distinct related constructs.

The purpose of language testing and the target language abilities being tested are essential to identify in the construction of a second language proficiency test and in the validation of the test score interpretation corresponding to differing proficiency levels. Given the frameworks of what of second language testing, the next section discusses several measurement methods employed in language tests and gives one example of a large-scale national English as a second language proficiency test and the measurement method employed in its score analysis.

1.2 English Language Proficiency Measurement

The key measurement issue in language tests is to determine the extent to which the sample of language use obtained from a test appropriately characterizes the overall potential language use of the individual. In the analysis of language test scores, several psychometric theories and models are employed, including observed score frameworks such as classical test theory (CTT) (Abedi, 2007) and generalizability theory (GT) (Bachman, 1997; Brown, 1999; Lynch & Mcnamara, 1998), and latent variable frameworks such as item response theory (IRT) (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; Brown,

1995; Sumita, Sugaya, & Yamamato, 2005), factor analysis, and structural equation modeling (SEM) (Kunnan, 1995). For example, Almond et al. (2005) used IRT models and the evidence centered design approach to analyze language proficiency as four language skill nodes: reading, writing, speaking, and listening. Several researchers also utilize testlet-based IRT and multidimensional IRT models in language tests (Min & He, 2014; So, 2010; Wainer & Wang, 2000).

Lately, many national and international English as a second language proficiency tests include listening, speaking, and writing components in addition to reading comprehension component. Some examples are the international Test of English as a Foreign Language (TOEFL), the International English Language Testing Service (IELTS), and national Test of English Proficiency (TEPS) in South Korea. Eignor, Golub-Smith, and Wingersky (1986) analyzed the fit of three-parameter logistic IRT model to TOEFL data that had three sections in 1984 and 1985 as listening comprehension, structure and written expression, and reading comprehension and vocabulary. They concluded that the model was a good fit to the reading comprehension and vocabulary items, but not a good fit to the dialogue items in the listening section and structure items in the TOEFL data. Similarly, TEPS currently has grammar and vocabulary components in addition to listening, speaking, writing, and reading. However, the TEPS employed the communicative competence framework and had only vocabulary, grammar, listening, and reading components in 1999 (Choi, 1999) in a multiple-choice format. Choi (1999) utilized an IRT approach to analyze these components and compared IRT scores with CTT observed scores. He argued that IRT is a better scoring method to use rather than raw scores from CTT for more precise measurement results.

Many countries employ national foreign language tests in addition to tests of academic subjects to make decisions about individuals (Bachman, 1990). One example of a national large-scale English as a second language test is the English Language Proficiency Test (ELPT) in a Middle Eastern country which was launched in 2013. The ELPT is administered two times a year by a national organization in charge of preparing and administering

national tests such as academic success, second language, medical certification, military and governmental employment large-scale tests including judge and attorney general appointments. The ELPT is a paper-pencil test taken by a variety of examinees, including: undergraduate and graduate students applying to graduate programs in various disciplines, people applying to positions in government agencies, people currently working in government agencies and interested in promotion in their positions, and people interested in knowing how proficient they are in English as a second language.

Decisions based on ELPT scores are about the selection or placement of individuals into post-secondary and workforce settings. For example, the selection might include whether an individual can enter a(n) (under)graduate program considering their readiness for academic instruction. The placement includes assigning employees into one of five categories of language proficiency with differences in compensation reflecting the potential contributions to their job using their knowledge in English. Occupations where the ELPT is used for placement include teachers in schools, instructors in academic institutions, judges, government appointed attorneys, translators, officers in the Ministry of Foreign Affairs, and department heads in government institutions. ELPT scores expire after five years, and the test must be re-taken or the proficiency level score decreases 10%. Thus, the ELPT is a paid high-stakes test for graduate school admission decisions and for determining salaries of government employees. ELPT scores affect graduate studies and career paths. The ELPT has different cut-off scores for different programs, job positions, and promotion depending on various institutional criteria. For example, the Graduate School of Natural and Applied Sciences in a university requires masters study applicants to have a minimum score of 50% and doctoral study applicants to have a minimum score of 55% in the ELPT. The Graduate School of Natural and Applied Sciences in another university requires masters study with a thesis in English applicants to have a minimum score of 55% in the ELPT whereas doctoral study with a thesis in English applicants to have a minimum score of 60%.

The ELPT is administered two times (Spring and Fall) a year with an approximate num-

ber of 450,000 test-takers per year. Unlike the large-scale second language proficiency tests examples mentioned above, the ELPT does not have writing, listening, speaking components; it tests reading ability, including items reflecting grammar, and vocabulary knowledge. The ELPT assesses test-takers English language proficiency with 80 five-option multiple-choice items in 180 minutes. Test-takers choose the best answer from among five alternatives based on their English language proficiency. Shuffling the items and their five-alternatives for security reasons (e.g. answer copying) result in multiple booklets delivered in paper-pencil format. Responses from the multiple booklets are mapped onto a single booklet and matched to the pre-determined answer key to score correct, incorrect, missing answers. The total number of correct responses is the ELPT score for each test-taker. There is no penalty to test-takers for getting a question incorrect, which means that at least making a guess for every item is the most effective way to proceed. Test scores are reported as the percentage of correct responses. The five-level proficiency in letters A, B, C, D, E are assigned to percentages as 90-100%, 80-89%, 70-79%, 60-69%, and 50-59% respectively.

The test has vocabulary, grammar, reading passages, sentence completion, translation, reading comprehension, dialogue, paraphrasing, and sentence omission items. Vocabulary items are sentences with a blank space asking for the best fitting word. Grammar items are sentences with one or two blank spaces asking for verbs in relevant tenses or prepositions, coordinating conjunctions, subordinating conjunctions, correlative conjunctions, and conjunctive adverbs. Several reading passages are paragraphs with blank spaces asking for conjunction, preposition, adverbs, or infinitives in the relevant tenses. Other reading passage items are paragraphs with a blank space, asking the examinee to choose the most relevant of a set of sentences to complete the paragraph. Sentence completion items present half of a sentence, asking the examinee to select the other relevant half. Translation items can ask for the translation of a sentence in either direction between English and the native language. Reading comprehension items are reading passages asking about

the inferences, main ideas, examples, and arguments in the passage. Dialogue items are printed-conversations missing one line asking for a relevant sentence. Paraphrasing items are sentences asking for the best relevant paraphrase. Lastly, sentence omission items are paragraphs with an irrelevant sentence to be omitted by the examinee.

In the reading comprehension, inferences involve a full range of organizational characteristics in the communicative competence framework. Paragraph reorganizing, vocabulary, and grammatical items also reflect organizational competence. In sentence omission and paraphrasing items, both organizational competence and pragmatic competence appear. In dialogue items, strategic competence is primarily dominant given the interchange between the context and discourse. In translation items, again strategic competence is required given there are two-way translations. These 80 items in the ELPT are related to English language ability primarily through reading comprehension component with a variety of five-alternative multiple-choice question formats. It was pointed out in the earlier literature on language tests that paper-pencil tests do not necessarily assess second language learners actual performance in communicative situations (Carroll, 1961; Oller, 1976). Given that the ELPT is not a computer-based test, the opportunity to assess speaking, listening, and writing components in the test is limited. The ELPT in its test structure leans primarily on English language reading ability of the test-takers. Thus, it may be suggested that the data may tend toward unidimensionality since writing, speaking, and listening are not measured as language domains in the ELPT.

To evaluate valid uses of this summative English language proficiency tests, this paper discusses what information IRT modeling provides over and above the CTT approach to the ELPT's scoring. The purpose of this study is to investigate whether an IRT model fits the ELPT data and serves as a more informative alternative to the current total number-correct scoring approach with dichotomously scored items in the area of language assessment. To address these issues, the study aims to i) compare the CTT and IRT approaches to scoring the ELPT, ii) explore person- and item-fit of the best fitting IRT model, iii) investigate

scores with large total score vs theta discrepancy, and iv) use multilevel modeling to analyze clustering of theta scores and omitting behavior. The next chapter describes the ELPT data and IRT models utilized in the psychometric analysis of its scoring.

CHAPTER 2

IRT Approach to Scoring the ELPT

This chapter describes the IRT modeling procedures used to analyze the ELPT data. The first section briefly describes how the data were coded and scored for IRT analysis. The next section presents the assumptions of IRT and describes the models applied to the data using the 'mirt' package in R environment (Chalmers, 2012). 'Mirt' is an open source package used for multidimensional item response theory analysis. The third section details the statistical approaches used to compare models in the IRT framework. Finally, the chapter details some of the advantages of using these IRT models for the ELPT data analysis.

2.1 Data

The data analyzed in the present study are the item-level responses of 288003 test-takers of the ELPT administered in Spring 2013. Several demographic variables such as gender and age were also in the data. Age was recorded in years and gender was treated as a dichotomous variable (0 for males, 1 for females). The mean of the age is 28 in an interval of [15, 73] and 48% of the test-takers are female. The answers to the items in the ELPT were coded 1, 0, and NA for correct, incorrect, and missing, respectively. Additionally, the data has regional testing centers coded in three digits, testing rooms coded in nine digits, and seat numbers coded in three digits where the ELPT was administered.

2.2 IRT Modeling

The overarching goal of this study is to utilize IRT analysis to estimate the test-takers language proficiency. The modeling process begins with stating the IRT assumptions. Then, it describes the unidimensional one-, two-, and three-parameter IRT models applied to the ELPT data.

An item response is often dichotomous in ability testing, i.e. getting the item correct or

incorrect. For such dichotomous responses, IRT assumes that the probability of answering an item correctly increases as the test-takers level of the latent trait increases. The IRT models have three main assumptions: unidimensionality, local independence, and monotonicity. Unidimensionality is the assumption that one dominant factor is measured by the items. Local independence means that after accounting for abilitys effect on test performance, any pair of the items in the test are statistically independent. Monotonicity is best displayed on a graph as an 'S'-shaped curve that is assumed to describe the relationship between the latent trait level on the x-axis and the probability of a response on the item on the y-axis. The monotonocity assumption is that the probability of a correct answer will increase as ability increases.

IRT models are in the form of a logistic function regressing the probability of correctly answering an item on the latent trait. The logistic model assumes that the logit of the probability of a correct response is linear in theta. The logit in Equation 2.1, where p is the probability of a correct response given the respondents ability, transforms the probability of a correct response from the interval [0, 1] to $[-\infty, +\infty]$ on the real number line.

$$logit(p) = log(\frac{p}{1-p}) \tag{2.1}$$

Considering Item Response Functions (IRFs), there are three main characteristics of the Item Characteristic Curves (ICCs) governed by three parameters: i) difficulty parameter, ii) item discrimination parameter, and iii) lower asymptote (pseudo-chance-level) parameter. The difficulty parameter of item i, labeled b_i, is the latent trait value at which a respondent is equally likely to answer correctly as to answer incorrectly. Higher values of b_i parameter require higher ability for a respondent to have a 50% chance to get the item correct, so the item is more difficult. The item discrimination parameter, labeled a_i, is proportional to the slope of the ICC at the point b_i on the ability scale, i.e. the steepness of the rise of the probability curve. Items with steeper slopes separate the respondents into different ability levels better than items with less steep slopes. The pseudo-chance-level parameter,

labeled c_i, is the lower asymptote for the item characteristic curve, i.e. the probability of respondents with low ability answering the item correctly.

We evaluate three unidimensional IRT models appropriate for dichotomous item response data that use one-, two-, or three-parameters mentioned above. Each model involves assumptions about the data (Hambleton, Swaminathan, & Rogers, 1991).

In the one-parameter IRT model often called the Rasch model (Rasch, 1960), the ICCs are given by the Equation2.2

$$P_i(Y = 1|\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}} \tag{2.2}$$

where $P_i(\theta)$ is the probability that any examinee with ability θ answers item i correctly and b_i is the difficulty parameter for item i, e is the transcendental number, and n is the number of items in the test. In this model, the only item characteristic affecting the respondents probability of a correct response is the item difficulty. It is assumed all items are equally discriminating. The lower asymptote is also zero, i.e. it is not allowed for low-ability respondents might guess.

In the two-parameter IRT model (Lord, 1952), the ICCs are given by the Equation 2.3

$$P_i(Y = 1|\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \tag{2.3}$$

where $P_i(\theta)$ and b_i are defined as in Equation 2.2. The a_i parameter is the item discrimination parameter for item i. In this model, in addition to item difficulty, item discrimination, proportional to the slope of the ICC at θ, is useful for discriminating the respondents near an ability level θ, i.e. separating respondents with abilities equal to or lower than θ from respondents with abilities greater than θ. The lower asymptote is zero, so guessing behavior is not considered in two-parameter model.

In three-parameter model (Birnbaum, 1968), the mathematical expression for the ICCs

is

$$P_i(Y=1|\theta) = c_i + (1-c_i)\frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}} \tag{2.4}$$

where c_i is the chance level parameter which represents the probability of an examinee with low ability answering the item correctly. This model reflects performance an at the lower end of the ability continuum where guessing is a factor, especially in multiple choice testing situations. Under the 3-PL model, the c-parameter affects the point of inflection of the item characteristic curve such that the probability of a correct response is $(1+c_i)/2$ rather than 0.5. The c_i parameter models the fact that low ability respondents may randomly guess on the items given structured alternatives. The model allows for the possibility that even participants with very low latent trait values have a certain probability of correctly answering the item (e.g. even very low ability individuals have a chance of guessing the correct response on multiple choice exams). The c_i parameter is theoretically between 0 and 1 but typically less than 0.3.

The item information functions denoted $I_i(\theta)$ can be calculated in the IRT framework. In the three-parameter model, the item information function is

$$I_i(\theta) = \frac{a_i^2(1-c_i)}{[c_i + e^{a_i(\theta-b_i)}][1+e^{-a_i(\theta-b_i)}]^2} \tag{2.5}$$

Information is higher when the b-parameter is close to the theta value, when the a-parameter is high, and as the c-parameter goes to zero. It describes the contribution of the items to the estimation of ability depending on the items discrimination power. If the c-parameter is greater than zero, the maximum information of an item is at an ability level slightly above its difficulty. The information function for a test at theta is the sum of all item information functions at theta as in Equation 2.6.

$$I(\theta) = \sum I_i(\theta) \tag{2.6}$$

It is possible to determine individual contribution of test items without knowing the other items in a test. The test information at theta is inversely related to the precision of ability estimation given by

$$SE(\overline{\theta}) = \frac{1}{\sqrt{I(\theta)}} \tag{2.7}$$

where $SE(\overline{\theta})$ is the standard error of estimation and it varies with ability level (Hambleton et al., 1991). The value of standard error depends on the number of test items, the quality of test items, the match between item difficulty and respondents ability level. Longer tests tend to have smaller standard errors. Highly discriminating items tend to result in smaller standard errors. Instead of having relatively easy or difficult tests, tests composed of items with difficulty parameters close to match the respondent ability tend to have smaller standard errors.

The 3-PL IRT model is the IRT model with all free parameters of item difficulty, item discrimination, and c-parameter. Restricting the c-parameter to zero, the reduced IRT model is the 2-PL model. Restricting all the item discriminations to one yields the Rasch model only with the item difficulty, which assumes all the items have the same slope or discrimination power. In this study, for all models, the item parameters were estimated with Bayesian estimation using 'mirt' package in R. The computation of IRT scale scores utilized the expected a posteriori (EAP) method (Bock & Mislevy, 1982).

2.3 Model Comparison

In the one-parameter model, a respondent's total score on the test is a sufficient statistic for estimating their latent trait. The two-parameter model usually provides a better fit to the data and adding the discrimination parameter allows items to vary in how they correlate with the trait. In ability tests, adding a third parameter is useful especially when items are multiple choice, i.e. possible guessing behavior from the structured alternatives. Given the multiple-choice structure of the ELPT items, the three-parameter model also has the

potential to model and explain the variation in the ELPT data. In the applications of IRT models, inadequacy of model-data fit may have adverse consequences such as biased ability estimates and unfair ranks (Wainer & Thissen, 1987; Yen, 1981). Thus, finding the best fitting IRT model to the ELPT data requires a comparison of specified models. This section details the statistical approaches used to compare models in the IRT framework.

When unidimensional IRT models are nested, comparison of the models can be done using likelihood ratio (LR) test to select relatively best fitting model (Kang & Cohen, 2007). The LR test statistic is calculated as the difference of the deviances from two compared models. This difference has a chi-square distribution and so significance tests can be employed to determine which model is the better fit (Baker & Kim, 2004). When IRT models are not nested, another approach for model selection is to use information statistics such as Akaike's information criterion (AIC) (Akaike, 1974), Schwarz's Bayesian information criterion or the deviance information criterion (DIC) (Spiegelhalter, Best, & Carlin, 1998) since they provide estimates for the relative differences between the model solutions, i.e. comparisons are based on relative magnitude. However, these statistics do not have related significance tests.

Models with more parameters generally tend to fit a data set better, especially if the models are nested (Kang & Cohen, 2007). To determine the best fitting model efficiently, using a significance test based on LR test is useful in such nested structures. However, AIC and BIC indices value the principle of parsimony and penalize the model for increasing the parameters as it gets more complex. The AIC is given by

$$AIC = -2log(likelihood) + 2p \tag{2.8}$$

where $-2log(likelihood)$ is the deviance statistic, and p is the number of estimated parameters. Better fitting models have smaller deviance, so AIC incorporates 2p as a penalty for overparametrization. The model with the smallest AIC is selected as the best fitting the

model. However, since it does not incorporate the sample size in its mathematical model, it is not asymptotically consistent (Schwarz, 1978). Using large data sample sizes, AIC tends to favor saturated models in very large samples (Janssen & De Boeck, 1999).

Schwarz (1978) suggested the BIC defined as

$$BIC = -2log(likelihood) + plog(N) \tag{2.9}$$

where N is the sample size. BIC penalizes overparameterization by incorporating a logarithmic function of the sample size in its calculation and so succeeds in finding asymptotic consistency. With large sample sizes, BIC tends to prefer simpler models than the ones selected by AIC. Since BIC penalizes the number of parameters at a higher level than AIC with large sample sizes, it is likely to favor simpler models with fewer parameters than AIC. Thus, these two statistics do not have to agree with each other.

Another index, DIC, deals with Bayesian posterior estimates of model parameters (Spiegelhalter, Best, Carlin, & Van der Linde, 2002). DIC is composed of a Bayesian measure of fit, the posterior mean deviance $\overline{D(\theta)}$ and a model complexity penalty, the number of free parameters in the model, p_d.

$$DIC = \overline{D(\theta)} + p_d = D(\overline{\theta}) + 2p_d \tag{2.10}$$

where $D(\theta)$ is the deviance at the posterior estimates of the parameters of interest and $p_d = \overline{D(\theta)} - D(\overline{\theta})$. The model with the smallest DIC is the model that would best predict a replicate data set with the same structure as the current observed.

IRT is a widely used approach to measurement moving beyond the constraints of CTT (Choi, 1999). It allows us to separate item and latent trait estimates, place them on the same scale, generate standard errors of estimation depending on the theta (ability) value in the unidimensional latent trait continuum, understand the contribution of each item to the standard error of estimation, and evaluate the fit of the specified models. Therefore,

IRT models provide information beyond the constraints in CTT such as test dependent true scores, sample dependent item characteristics and reliability estimates, constant standard error of measurement for all test-takers, each item given equal weight, and limited number of observed scores. Given a background on the IRT models utilized in the analysis of the ELPT in this chapter, the next chapter details the results of the IRT and CTT analysis using the ELPT data.

CHAPTER 3

Fitting a Unidimensional IRT Model

This chapter fits the language test data using both a classical test theory approach and a unidimensional IRT model. The best fitting unidimensional IRT model is the 3-PL, for which I discuss the parameter estimates, test information function, the person and item statistics.

3.1 Classical Test Theory

The reliability coefficient Cronbachs alpha for the ELPT is 0.93, which shows that the ELPT proves to be quite reliable from the perspective of CTT. The mean of total scores is 30.46 out of 80 (SD=14.72) in Table 3.1. The standard error of measurement (SEM), therefore, is 3.9 out of 80 points, which was calculated by Equation 3.1.

$$SEM = SD\sqrt{1-\alpha} \tag{3.1}$$

where α is the Cronbachs alpha and SD is the standard deviation of the total observed scores.

As shown in Figure 3.1, the distribution of the total scores is positively skewed. Additionally, total number-correct scores of the test-takers significantly correlate with the age, r=0.16. Total score increases as examinee age increases. Men (M=30.72, SD=14.92) reported significantly higher total scores than women (M=30.19, SD=14.49), $t(287180)$=9.68.

Table 3.1: *Descriptive statistics of answers*

Answers	Mean	*SD*
Correct	30.46	14.72
Incorrect	47.97	14.99
Missing	1.57	6.79

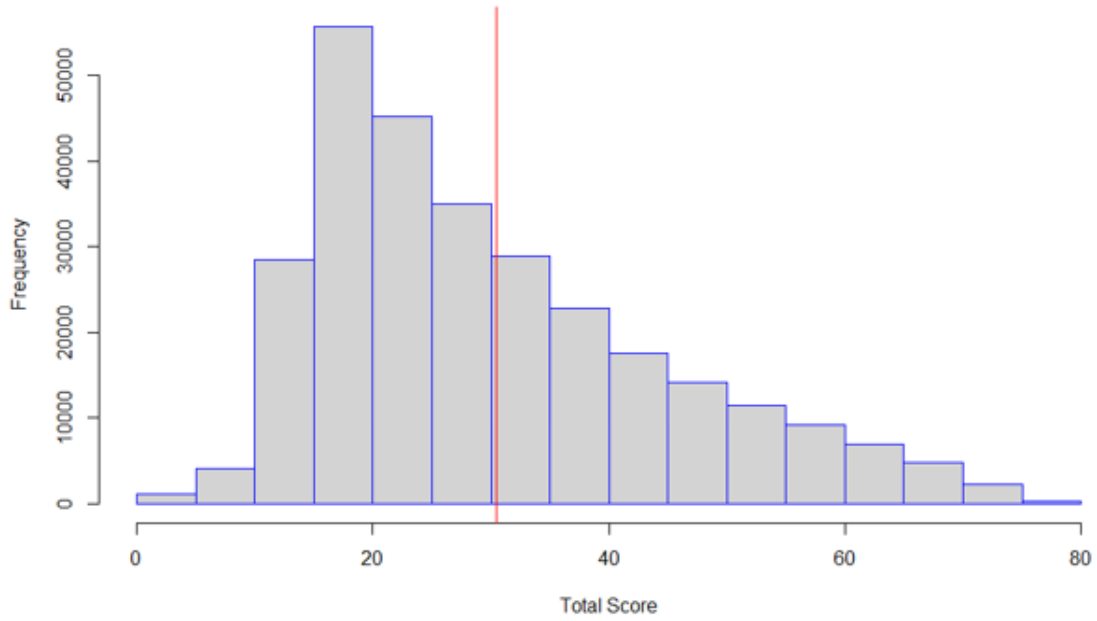


Figure 3.1: Total score distribution

Table 3.2 presents the item discrimination and difficulty parameters under classical test theory. Ebel and Frisbie (1986) suggested that good items have a discrimination index of 0.40 and higher; reasonably good items from 0.30 to 0.39; marginal items from 0.20 to 0.29, and poor items less than 0.20. For the ELPT, the overall discrimination power obtained by the mean item-total correlation is 0.37, which is over the criterion 0.29, so the test by and large succeeds in discriminating well among the test-takers. However, items 36 and 56 appear to have very low discrimination power as 0.02 and 0.07, respectively. That means these items are not good enough to distinguish the high language ability and low language ability intended to be measured by the ELPT.

A way to calculate the ideal difficulty level is to identify the point on the difficulty scale midway between perfect, i.e. 100% of examinees, and chance-level difficulty, i.e. 20% of examinees for items with five alternatives. The optimal difficulty level for 5-alternative items is, thus, 60% (Thompson & Levitov, 1985). For the ELPT, the item difficulty index

of mean proportion correct is 0.39 presented in Table 3.2, suggesting that the test proves to be more difficult than expected optimal difficulty level. However, items 36 and 56 again appeared noticeably to deviate from the mean proportion correct. The difficulty parameters for item 36 and 56 are 0.21 and 0.14, respectively and so they do not match the language ability of most of the test-takers given these might be unnecessarily difficult questions.

Item analysis based on CTT indicates that the mean of the difficulty and discrimination parameters are 0.37 and 0.39, respectively. Considering the large sample size, these item statistics provide a basis for items characteristics such as potential problematic items. Given the item analysis based on this sample, it is not possible to compare the test-takers scores in this ELPT to the scores of other test-takers in the ELPTs administered different year or seasons, which may cause unfair assignments to the ability category levels in different administrations.

Additionally, CTT-based descriptive statistics provided insight into the magnitude of standard error of 3.9 which is constant for every test-taker. However, since the ELPT has a scoring approach based on only the number of correct answers, it might be misleading to have the same standard error for every test-takers score given they might have different numbers of missing answers, which implies that they have taken different tests, not the same 80-item test, so the standard error may not be the same for every test-taker.

The ELPT treats the missing answers as if they were wrong answers given the number-correct scoring approach. However, it has been emphasized by Lord (1974) that missing answers should not be ignored or treated incorrect since test-takers may not be using the same test-taking strategy, which adds another factor on total score beyond the intended construct. In CTT, there are no methods to handle the missing data and integrate it into scoring to better reflect the ability differences whereas IRT algorithms handle missing data well.

In addition, the ELPT is a multiple-choice format test giving the test-takers an opportunity to randomly guess answers to the items. The test-takers are also not penalized in their

Table 3.2: *Item discrimination and difficulty under CTT*

Item	*n*	*p*	*r*	Item	*n*	*p*	*r*
Item 1	284971	0.53	0.45	Item 41	286785	0.47	0.34
Item 2	282607	0.35	0.38	Item 42	286794	0.64	0.49
Item 3	284407	0.35	0.39	Item 43	280586	0.47	0.43
Item 4	283158	0.19	0.36	Item 44	278047	0.33	0.45
Item 5	283816	0.29	0.33	Item 45	279241	0.28	0.27
Item 6	284853	0.49	0.56	Item 46	277715	0.50	0.44
Item 7	285315	0.52	0.44	Item 47	280620	0.37	0.31
Item 8	286344	0.50	0.46	Item 48	278412	0.27	0.17
Item 9	284158	0.45	0.35	Item 49	277886	0.41	0.26
Item 10	285740	0.36	0.47	Item 50	278600	0.36	0.38
Item 11	286026	0.38	0.18	Item 51	280041	0.26	0.30
Item 12	285246	0.51	0.47	Item 52	279914	0.18	0.36
Item 13	284928	0.34	0.40	Item 53	277692	0.30	0.40
Item 14	285537	0.45	0.58	Item 54	277647	0.34	0.40
Item 15	285752	0.45	0.44	Item 55	280181	0.33	0.28
Item 16	284351	0.41	0.35	Item 56	279677	0.14	0.07
Item 17	285743	0.46	0.15	Item 57	279273	0.31	0.34
Item 18	285718	0.27	0.33	Item 58	278343	0.34	0.44
Item 19	285507	0.18	0.38	Item 59	279319	0.32	0.30
Item 20	283156	0.36	0.23	Item 60	280212	0.40	0.47
Item 21	284125	0.50	0.33	Item 61	279402	0.45	0.39
Item 22	284525	0.23	0.40	Item 62	278405	0.37	0.57
Item 23	283583	0.32	0.29	Item 63	283307	0.41	0.43
Item 24	281315	0.46	0.22	Item 64	285160	0.38	0.41
Item 25	283724	0.42	0.41	Item 65	285516	0.65	0.55
Item 26	284921	0.27	0.35	Item 66	282882	0.42	0.38
Item 27	280418	0.34	0.39	Item 67	283852	0.46	0.39
Item 28	284522	0.38	0.31	Item 68	278354	0.44	0.46
Item 29	284175	0.28	0.41	Item 69	279012	0.37	0.30
Item 30	282343	0.38	0.35	Item 70	278637	0.39	0.51
Item 31	283846	0.42	0.47	Item 71	280015	0.42	0.39
Item 32	283287	0.32	0.45	Item 72	275405	0.30	0.14
Item 33	281627	0.33	0.28	Item 73	277981	0.32	0.32
Item 34	284442	0.56	0.49	Item 74	278628	0.25	0.38
Item 35	282456	0.19	0.35	Item 75	277822	0.33	0.51
Item 36	282865	0.21	0.02	Item 76	281430	0.37	0.32
Item 37	286752	0.64	0.46	Item 77	282196	0.59	0.41
Item 38	286074	0.56	0.39	Item 78	282882	0.48	0.46
Item 39	286792	0.40	0.37	Item 79	282745	0.47	0.49
Item 40	287084	0.72	0.40	Item 80	282293	0.38	0.16

n : number of responses, *p* : item difficulty, *r* : item discrimination

scores for their random guessing behavior on the items where they do not know the answer or do not have enough time. IRT models allow sample independent item analysis and individual ability estimates, provide individual standard errors for more precise measurement information, handle the missing data, and allow for modeling guessing behavior. Thus, the use of unidimensional IRT models would be informative to go beyond the constraints of the CTT-based scoring approach and provide more insight about the items, scores, and the test characteristics.

3.2 Fitting the 3-PL IRT Model

The goal is to fit the most appropriate IRT model to explain the ELPT data with a better scoring method for the second language ability independent of the constraints of total number-correct scoring approach in the CTT framework. As opposed to the CTT, an IRT model would handle the missing data and provide further information about the items, scores, and the test independent of the current sample such as item parameters and individual standard errors of ability. Thus, this section details the IRT model comparison to identify the best fitting model to the ELPT data, and then presents the item parameter estimates, test information function, person- and item-fit statistics given the best fitting the 3-PL model.

3.2.1 Unidimensional IRT Model Comparison

For the analysis of the ELPT in the theoretical framework of item response theory, I applied three unidimensional IRT models to dichotomously scored data using the 'mirt' package in R. The three unidimensional IRT models applied were 1-parameter logistic (Rasch), 2-parameter logistic (2-PL), and 3-parameter logistic (3-PL) model.

An initial analysis with maximum likelihood estimates of model parameters was done by running all three models. The likelihood ratio test statistic was employed given the nested structure of the models. AIC and BIC compared their relative magnitude of model solutions. In very large samples as in the ELPT, the AIC tends to select saturated models

Table 3.3: *Unidimensional Model Fit Statistics*

Model	*DIC*	*AIC*	*BIC*	LR Test
1-PL (Rasch)	26400856	26400584	26401441	
2-PL	26032438	26032035	26033727	368707.1*
3-PL	25850461	25852821	25855358	179374.6*

$*p < .001$

(Janssen & De Boeck, 1999). The BIC tends to prefer simpler models than the ones preferred by the AIC when the sample size is large. The relative penalties for BIC and AIC are $plog(N)$ and $2p$, relatively, where p is the number of parameters and N is the sample size. The larger the sample sizes the larger the penalty in calculation of the BIC statistic, so the BIC prefers less number of parameters suggesting simpler models compared to the AIC. Table 3.3 shows the statistics of the models.

AIC, BIC, and LR gave consistent results in identifying the 3-PL as the best model for ELPT data: The smallest AIC was for the 3-PL and LR test indicated that the 3-PL model explained the data better than the 2-PL and that the 2-PL explained the data better than 1-PL (Rasch) model. The BIC was also the smallest for the 3-PL, pointing to the 3-PL as the best fitting model. Even though AIC and BIC do not necessarily agree with each other (Lin & Dayton, 1997), the two statistics agreed in the ELPT data and all three indices suggested the 3-PL is the best model for the ELPT data. However, several items were problematic such that the item discrimination estimate for item 36 was negative (a-parameter = -4.16) and item difficulty estimate for item 56 was too high (b-parameter = 56).

Therefore, a second analysis was run with a log normal prior on the a-parameters and a wide normal on the b-parameters for all three models. This increased the speed of the convergence and eliminated the bad item parameter estimates. Table 3.3 also shows the DICs for the models that deal with the Bayesian posterior estimates of the model parameters. The DIC was also the smallest for the 3-PL, pointing to the 3-PL as the best model that would predict a replicate data set of the same structure as the ELPT. Thus, all four methods

to compare the fit of the models determined the 3-PL as the best fitting model to the ELPT. Given the ELPT is a multiple-choice test, the test-takers have a potential to randomly guess the correct answers. This supports the unidimensional 3-PL model fitting to the data better than other two unidimensional IRT models since it accounts for the pseudo-chance-level parameter (c-parameter) and it provides additional information within this model explaining the variation in the data better.

3.2.2 3-PL Parameter Estimates

The 3-PL item parameter estimates are in Table 3.4 with the item loadings for the dimension of English language proficiency. All loadings are above 0.4, which is the recommended cut-off value for factor loadings on a dimension (McCoach, Gable, & Madura, 2013). The loadings for the one factor in the 3-PL model all supported a one factor solution with higher loadings then those of 1- and 2-PL, i.e. all item loadings are above 0.4.

The b-parameter estimates range from -0.93 to 3.13; item 40 is the easiest item and item 56 is the most difficult item. 76 of the b-parameters are above zero and only four of them are below zero, which suggests that most of the items in this test are relatively difficult and therefore most useful in addressing the ability levels of test-takers with higher than the average ability. The mean of the item difficulty is 1.03.

The a-parameter estimates range from 0.6 to 4.34 with a mean of 1.97. 47 items are below the mean discrimination power and 33 items are above it. Among 80 items in the test, the slope of the item characteristic curve for item 75 is the steepest and the item characteristic curve for item 24 has the flattest slope, i.e. it is barely discriminating the test-takers with high ability from the ones with low ability. A low discrimination index occurs for items that are too hard or poorly written, which makes it difficult to select the correct answer. On these items, low ability students may guess correctly, while more skilled students, suspecting that a question is too easy, may answer incorrectly by reading too much into the question. Or there may be a few other very plausible alternatives to the

answer of question and it is hard for test-takers with high ability to differentiate among those alternatives.

The c-parameter ranged from 0 to 0.39 and all items have c-parameters lower than 0.3 except for item 1, 17, and 21. The lower asymptotes of all items reflect non-zero performance of low proficiency examinees on multiple choice items except for items 16 that has a zero c-parameter. Additionally, items 37 and 40 have very low c-parameters of 0.01, meaning that examinees can answer this item correctly just by guessing with a very low probability of 1%.

3.2.3 Test Information Function

In the context of IRT, the amount of information that each item or test provides is not equal across the entire continuum of latent constructs. Given the parameter estimates and the item information curves for items 24 and 75, items with higher a-parameter values are more informative than items with lower a-parameter values. The information curve, for example, is flatter for item 24 than item 75 since the slope of item 24 is smaller than the slope of item 75. For the individual items, the dichotomous response item provides most of the information accumulating around the b-parameter which is shown by the vertical lines on the ICCs and IICs for item 24 and 75 in Figure 3.2.

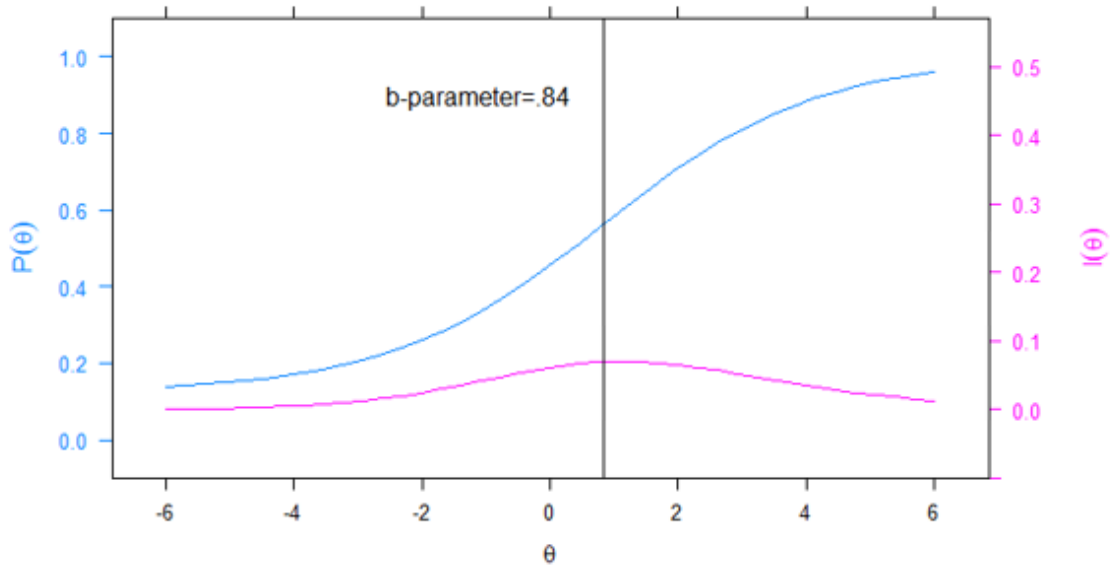
The test information curve for the unidimensional 3-PL model, in Figure 3.3, provides most of the information between theta scores of the interval [0.5, 2] in the theta range from -2.04 to 3.44. The sum of the information function curves for 80 items accumulates in [0.5, 2] the most and so provides most of the information about the test in this interval. Thus, for examinees with theta between 0.05 and 2, the test addresses their ability level with more precise measurement.

The test information at theta also enables us to construct a confidence interval for interpreting the ability estimate. Using the normality of ability distribution, we can construct a confidence interval for θ. For example, for a $(1-\alpha)\%$ confidence interval we

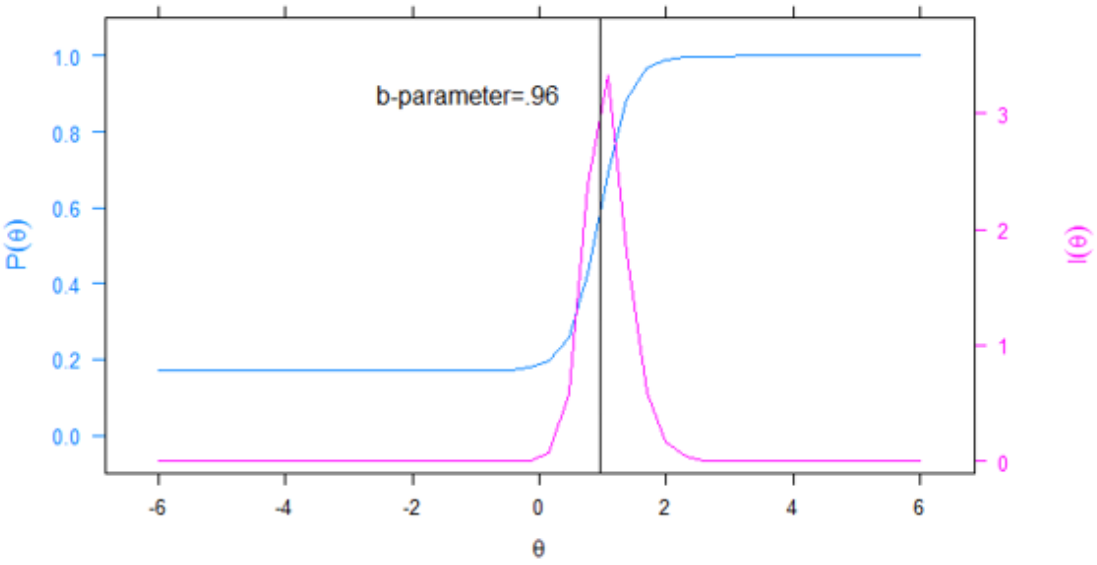
Table 3.4: *Item parameter estimates for unidimensional 3-PL IRT model and loadings*

Item	*a*	*b*	*c*	*L*	Item	*a*	*b*	*c*	*L*
Item 1	2.68	0.59	0.31	0.65	Item 41	1.05	0.64	0.15	0.53
Item 2	1.45	1.09	0.14	0.82	Item 42	2.08	-0.2	0.18	0.77
Item 3	2.46	1.16	0.21	0.89	Item 43	1.86	0.65	0.22	0.74
Item 4	3.4	1.51	0.11	0.75	Item 44	2.06	1.02	0.14	0.77
Item 5	1.9	1.45	0.17	0.81	Item 45	1.34	1.73	0.15	0.62
Item 6	2.32	0.26	0.11	0.61	Item 46	1.92	0.56	0.23	0.75
Item 7	1.31	0.09	0.07	0.69	Item 47	1.11	1.25	0.16	0.55
Item 8	1.6	0.3	0.14	0.53	Item 48	1.7	2.12	0.22	0.71
Item 9	1.05	0.61	0.13	0.86	Item 49	1.42	1.44	0.28	0.64
Item 10	2.9	0.97	0.18	0.63	Item 50	1.97	1.14	0.2	0.76
Item 11	1.39	1.92	0.3	0.73	Item 51	1.89	1.6	0.16	0.74
Item 12	1.79	0.32	0.16	0.76	Item 52	2.47	1.6	0.09	0.82
Item 13	1.96	1.13	0.17	0.79	Item 53	1.57	1.18	0.11	0.68
Item 14	2.2	0.33	0.08	0.70	Item 54	2.63	1.16	0.2	0.84
Item 15	1.69	0.61	0.17	0.48	Item 55	1.04	1.54	0.15	0.52
Item 16	0.92	0.46	0	0.63	Item 56	1.68	3.13	0.13	0.70
Item 17	1.38	1.96	0.39	0.80	Item 57	2.49	1.37	0.2	0.83
Item 18	2.27	1.48	0.17	0.87	Item 58	2.65	1.04	0.18	0.84
Item 19	2.99	1.53	0.1	0.89	Item 59	1.81	1.48	0.2	0.73
Item 20	3.4	1.6	0.3	0.69	Item 60	2.19	0.8	0.17	0.79
Item 21	1.61	0.91	0.31	0.78	Item 61	1.31	0.67	0.15	0.61
Item 22	2.12	1.41	0.1	0.62	Item 62	3.22	0.74	0.14	0.88
Item 23	1.36	1.51	0.18	0.33	Item 63	1.91	0.82	0.19	0.75
Item 24	0.6	0.84	0.12	0.68	Item 64	1.29	0.75	0.08	0.61
Item 25	1.57	0.8	0.17	0.79	Item 65	2.52	-0.42	0.05	0.83
Item 26	2.21	1.43	0.16	0.86	Item 66	1.64	0.94	0.21	0.69
Item 27	2.86	1.16	0.21	0.58	Item 67	1.57	0.7	0.21	0.68
Item 28	1.2	1.22	0.18	0.74	Item 68	2.22	0.72	0.2	0.79
Item 29	1.87	1.24	0.12	0.79	Item 69	2.46	1.35	0.26	0.82
Item 30	2.21	1.19	0.24	0.83	Item 70	2.83	0.83	0.18	0.86
Item 31	2.51	0.82	0.21	0.77	Item 71	1.66	0.89	0.2	0.70
Item 32	2.08	1.06	0.13	0.63	Item 72	1.26	2.45	0.24	0.59
Item 33	1.37	1.52	0.19	0.75	Item 73	3.34	1.37	0.23	0.89
Item 34	1.92	0.11	0.16	0.91	Item 74	3.53	1.36	0.15	0.90
Item 35	3.77	1.51	0.11	0.88	Item 75	4.34	0.96	0.17	0.93
Item 36	3.14	2.49	0.2	0.66	Item 76	0.91	1.02	0.09	0.47
Item 37	1.48	-0.51	0.01	0.67	Item 77	1.66	0.19	0.26	0.70
Item 38	1.54	0.39	0.26	0.58	Item 78	1.42	0.28	0.09	0.64
Item 39	1.2	0.85	0.13	0.62	Item 79	2.04	0.47	0.17	0.77
Item 40	1.35	-0.93	0.01	0.65	Item 80	0.78	2.27	0.25	0.42

a : a-parameter, *b* : b-parameter, *c* : c-parameter, *L* : loadings



(a) Item 24



(b) Item 75

Figure 3.2: ICCs and IICs

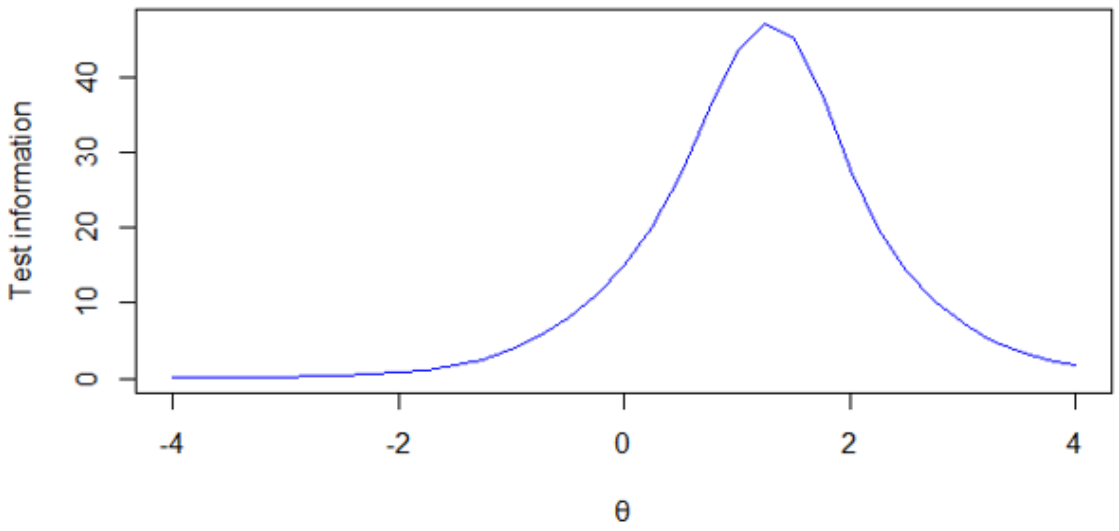


Figure 3.3: Test information function

obtain the confidence interval as $[\overline{\theta} - z_{\alpha/2} * SE(\overline{\theta}), \overline{\theta} + z_{\alpha/2} * SE(\overline{\theta})]$ where the standard error of $\overline{\theta}$ is $SE(\overline{\theta})$ and $z_{\alpha/2}$ is the percentile point of the normal distribution. That is, we are $(1-\alpha)\%$ confident that the real ability (θ) of a test-taker is in the interval $[\overline{\theta} - z_{\alpha/2} * SE(\overline{\theta}), \overline{\theta} + z_{\alpha/2} * SE(\overline{\theta})]$. In this regard, standard error of the ability estimate in IRT is similar to the standard error in CTT, but SE value varies with the ability level instead of having a constant SE for any test-taker. For example, in the ELPT context, the standard error of the ability estimates, in Figure 3.4, is much less for the test-takers with theta scores in the interval [0.5, 2]. The confidence bands for theta is much narrower for examinees with estimated ability in this interval, which provides a more precise measurement error.

The magnitude of the standard error depends on the number of test items, the quality of test items, the match between item difficulty and respondents ability level. The curve of the standard error of the ability in Figure 3.4 enables us to conclude that the items in the ELPT are highly discriminating especially for the test-takers whose ability level is in the interval [0.5, 2] since highly discriminating items tend to result in smaller standard errors. Additionally, the standard ability curve shows that the test has items with difficulty parameters close to match the test-taker ability in the interval [0.5, 2].

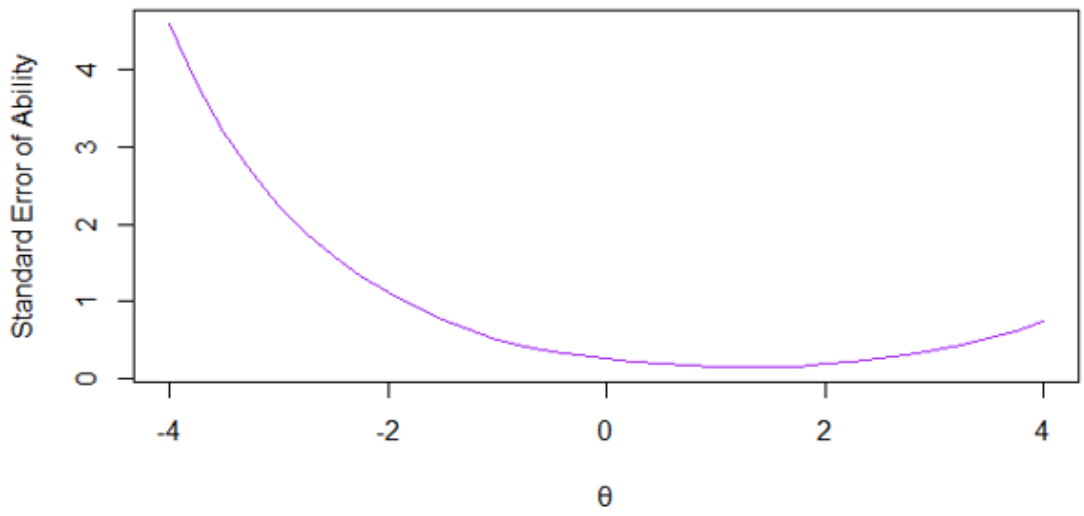


Figure 3.4: Standard error of ability

3.2.4 Person Fit statistics

The model fit can be evaluated at the test, item, and person level. Different statistical approaches exist to determine whether item score patterns provide additional information to the total score of the test. In person-fit analysis, an observed item score pattern on a test is compared to the expected item score pattern that is determined by an IRT model or the behavior of majority of test-takers (Meijer & Sijtsma, 2001). A large difference between the observed and expected pattern indicates misfit. For example, a test-taker answering difficult items correctly and easy ones incorrectly (or randomly guessing on all items due to time constraints) would produce a large person-misfit index value. They have aberrant response patterns since such patterns are unlikely to be observed based on the model. If such response patterns are unlikely then the answers might be driven by another mechanism other than the targeted construct (Meijer, 2003). Person-misfit might occur due to several reasons such as guessing, cheating, or different question-answering approaches.

The Zh statistic is a standardized person-fit value of l_z, which is used for categorical

data (Drasgow, Levine, & Williams, 1985). The Zh is defined as

$$Zh = \frac{l_z - E(l_z(\theta))}{SD(l_z(\theta))} \tag{3.2}$$

where $E(l_z(\theta))$ is the mean l_z-value for the sample, and $SD(l_z(\theta))$ is the standard deviation. This transformation standardizes the value to have a mean of 0 and a standard deviation of 1 by dividing the difference between l_z and mean l_z by the standard deviation of the observed value. The overall model and person fit were all estimated in the R programming environment with the multidimensional IRT package 'mirt' (Chalmers, 2012).

In the IRT context, when individual item parameter values are estimated within the model, then Zh distribution is formed. The distribution of Zh is typically non-normal, which makes it different than z-statistic distribution (Drasgow et al., 1985). Because of non-normality, the cut-off of -1.96 could be just a starting point to identify misfitting responses. A visual inspection of the distribution of the Zh-values in Figure 3.5, obtained from unidimensional 3-PL model, I chose to form a cut-off of 3 to serve as a reference point in identifying the aberrant answer patterns for the test-takers in the ELPT. Test-taker below or above this cut-off do not necessarily represent the outliers, but test-takers with unusual answer patterns suggesting further investigation.

With respect to this cut-off value, there are 1330 test-takers with person-fit scores below -3 and 21 test-takers above +3. 1088 of the 1330 test-takers have no omitting behavior, which suggest that there might be guessing behavior in their test-taking strategy. Additionally, 530 of 1330 had a total score higher than the average total score of 30. A larger cut-off value might be used to move from as the aberrancy of test-takers in the extreme comes clear.

An example of an atypical case is the response pattern of an examinee with a person fit score below -5. The examinee has a total score of 50 and a theta score of -0.56. The mean of b-parameter estimates of the items answered correctly is 1.18, but the mean of b-

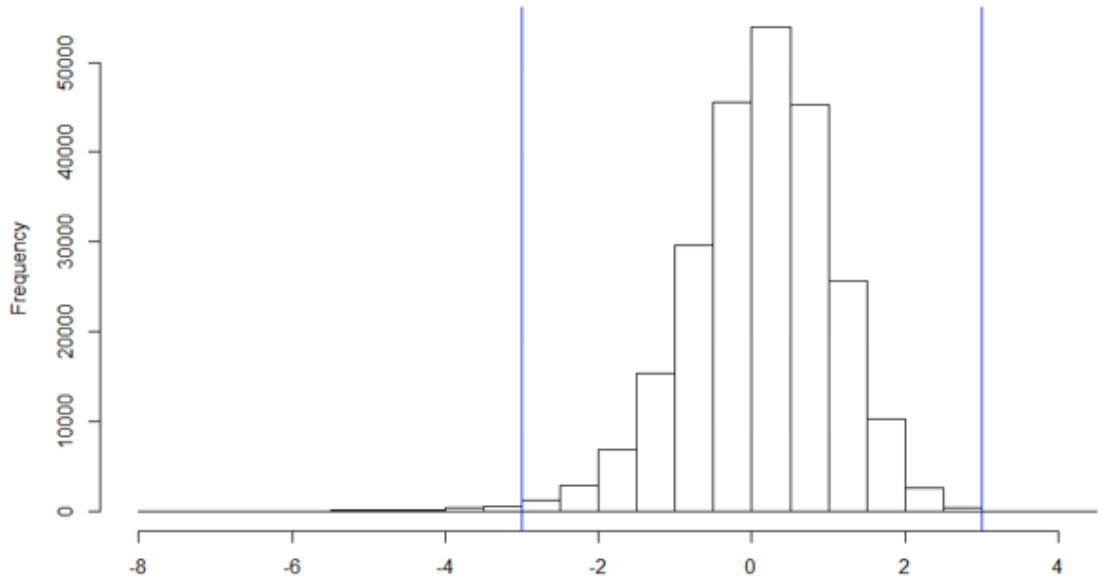


Figure 3.5: Person fit distribution

parameter estimates of the items answered incorrectly 0.77. Thus, s(he) answered most of the items with difficulty levels higher than the mean difficulty of all items (1.03) correctly but most of the items with difficulty levels lower than the mean difficulty of all items (1.03) incorrectly.

3.2.5 Item Fit Statistics

There are more statistical approaches developed to evaluate item-fit rather than overall model-fit (Embretson & Reise, 2000). One reason might be that even when the model overall fits the data, some of the items may not function in the intended manner. Inadequacy of model-data fit may have negative consequences in the application of IRT models such as biased ability estimates and unfair ranking (Wainer & Thissen, 1987). Item-fit analysis also enables test developers to identify bad items given the purpose of the test and retain only items fitting an IRT model in line with the measurement of the targeted construct. Assessing a model fit at the item level, thus, substantially contributes to model fit analysis. Item fit was also estimated with the multidimensional IRT package 'mirt' (Chalmers, 2012).

Orlando and Thissen's (2000) item fit statistic, $S-\chi^2$, for dichotomously scored items conditions on summed scores instead of ability. Observed and expected frequencies for each summed score are compared using chi-square statistic. Expected frequencies computation utilizes the joint likelihood distribution of each possible total score across all possible response patterns for each total score. In other words, expected frequencies are based on the likelihood of response patterns where a given item is answered and which produce a given total score.

Based on the simulations, Orlando and Thissen (2000) suggested that $S-\chi^2$ performs well with acceptable Type I error rates and large sample sizes further support adequate power. The $S-\chi^2$ statistic is preferable with several positive characteristics. Research indicated that it had acceptable Type I error rates and adequate power for large samples (e.g. larger than 2000) for dichotomous IRT models (Orlando & Thissen, 2003; Stone & Zhang, 2003). Additionally, $S-\chi^2$ does not divide theta into random intervals as it is conditioned on summed total scores. However, while the number of items or alternatives for items increases, the number of possible total scores increases, which causes sparsity. The histogram of the item fit statistic $S-\chi^2$ is shown in Figure 3.6.

One of the well fitting items in the test is item 5 with a small $S-\chi^2$ shown by in Figure 3.7a, where observed values nicely align with the expected values on the item characteristic curve.

However, the histogram of $S-\chi^2$ in Figure 3.6 also indicated that several items are misfitting with relatively large differences in expected and observed frequencies. For example, the empirical estimate for item 36 is plotted against the expected estimate from the model in Figure 3.7c. This is the item identified as problematic in both CTT and in the maximum likelihood estimation of the model parameters, which performed better after introducing the prior. The empirical plot based on item fit analysis prove the misfit between the model and the item 36 such that the item is still problematic with evidence of non-monotonicity. Similarly, item 56 is still problematic with evidence of non-monotonicity

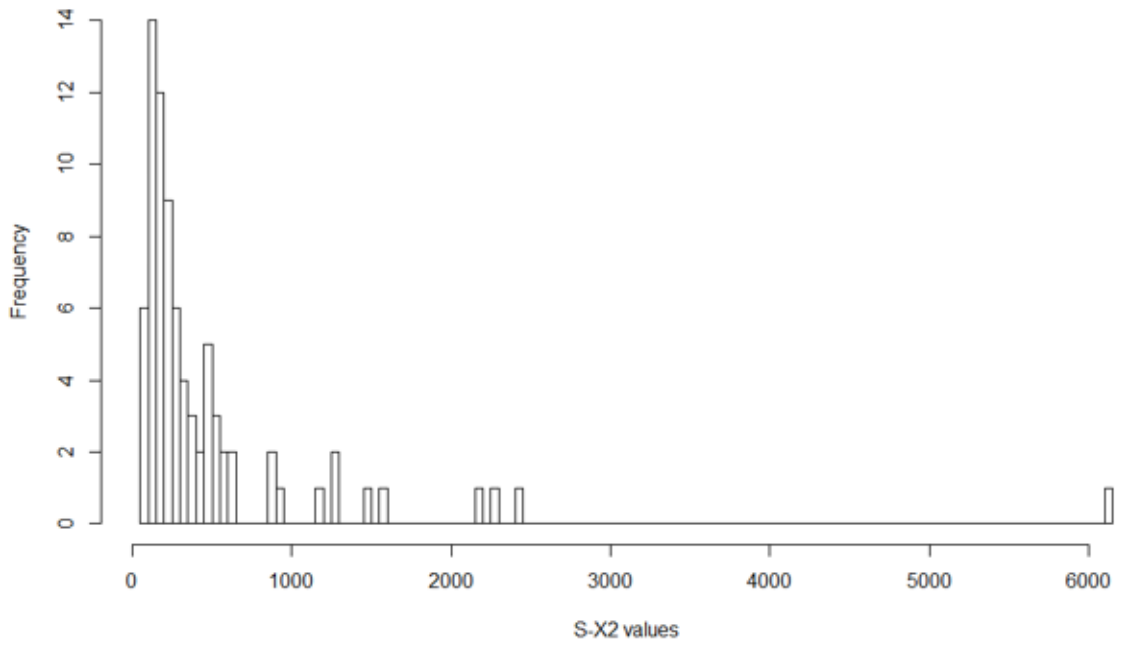
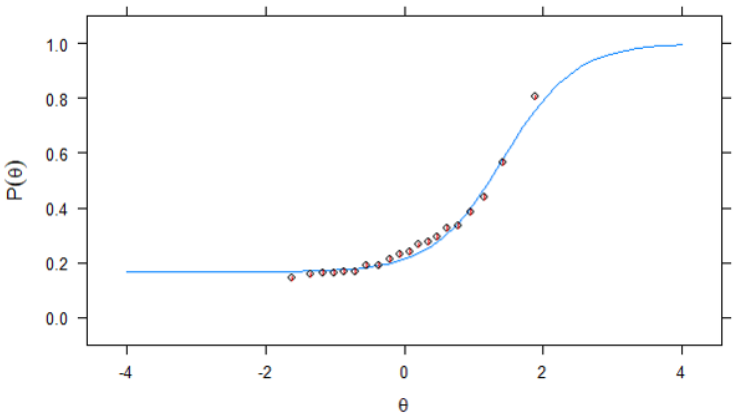


Figure 3.6: Item fit distribution

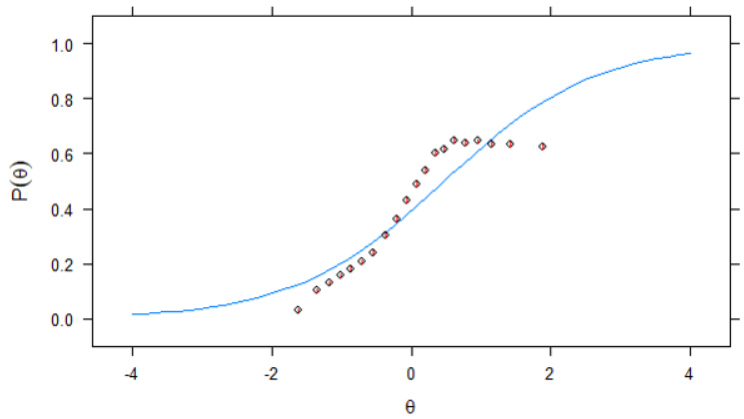
based on its empirical plot.

Based on the histogram of item fit statistic, item 16 was the worst fitting item with the highest $S-\chi^2$ value. The empirical plot of item 16 in Figure 3.7b shows that this item is also problematic with evidence of non-monotonicity. In CTT item analysis, item 16 was not flagged, but the item fit analysis in IRT was able to capture that this item also needs further attention. Thus, item fit analysis in IRT can easily point out the items with problems and in need of further development in the test.

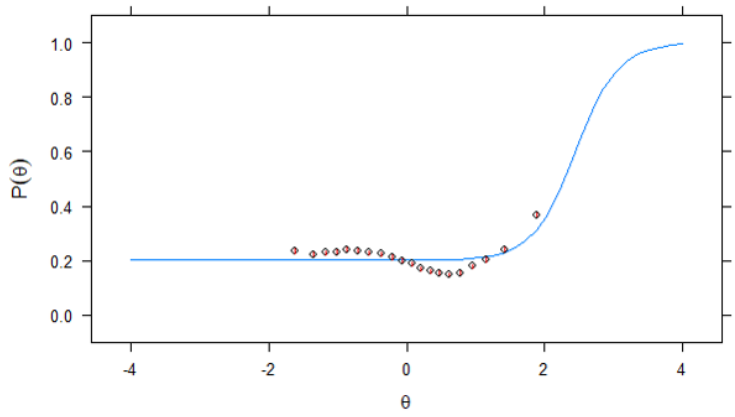
The analyses in this chapter indicated that CTT provides a sample dependent item analysis, a constant standard error of the measurement for all examinees, and a reliability estimate based on parallel forms. The 3-PL IRT model provides more information about item parameter estimates, ability estimates, item and test information, person and item fit. Including a pseudo-chance-level c-parameter with the 3-PL model accounts for possible guessing behavior given that the ELPT has a multiple-choice structure. Sample independent parameter and ability estimation algorithms could handle the missing answers in the



(a) Item 5



(b) Item 16



(c) Item 36

Figure 3.7: Empirical plots

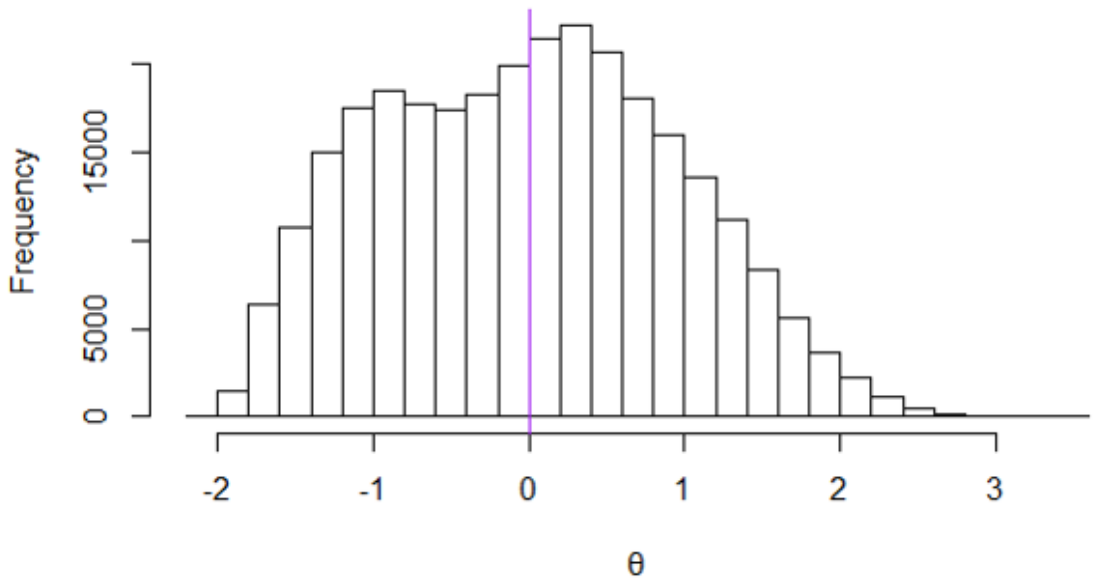


Figure 3.8: Theta distribution

ELPT data. Additionally, the misfit between examinee response pattern and their language ability could be estimated through several person and item fit indices. Item information functions depict how each item contributes to the overall test information function, which estimates the measurement precision at the individual level. In CTT, every item is assumed to contribute to the measurement of the intended construct at the same amount, but item information functions with item parameter estimates based on IRT models also might help deciding whether to retain items for further use or drop some items because they do not significantly contribute to the measurement of the intended latent construct.

Since total number correct score and theta score are two different scoring approaches to the same ELPT data, the relationship between the two is also important to further explore. Estimated theta scores can be compared with the total scores. The estimated ability score distribution obtained from the IRT model, in Figure 3.8, is bimodal as opposed to the positively skewed distribution of total scores. The bimodal structure may imply an underlying two factor solution as evidence of multidimensionality in the data.

Theta scores of the test-takers significantly correlate with the age, r=0.16. Theta score

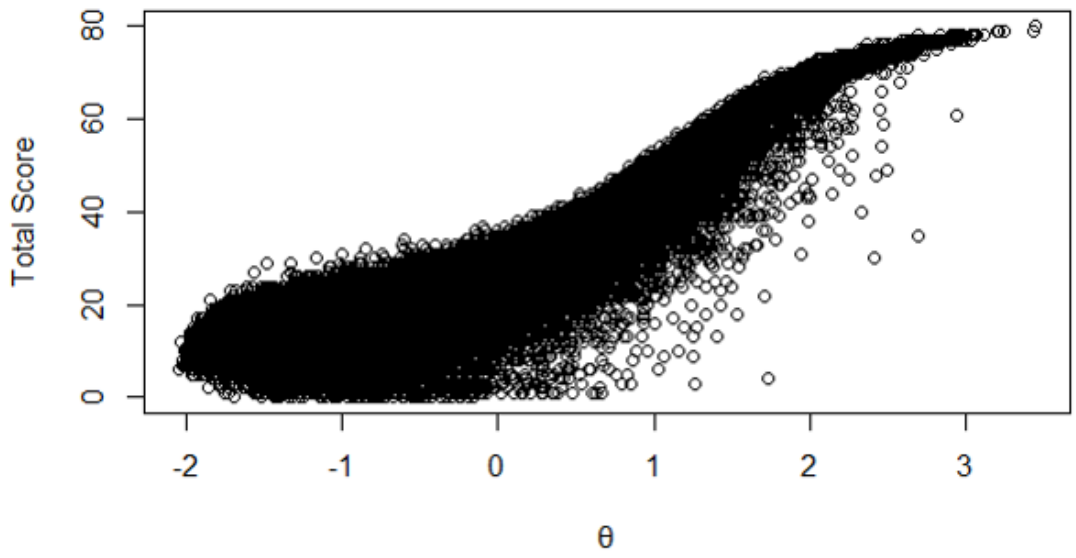


Figure 3.9: Total score vs theta

increases as examinee age increases, which is also the case in total scores. However, theta scores from the 3-PL model suggested that there is no statistically significant difference between women and men in their theta scores, $t(287210)=0.63$, $p=0.53$ whereas men ($M=30.72$, $SD=14.92$) recorded significantly higher total scores than women ($M=30.19$, $SD=14.49$). Thus, theta scores did not suggest any gender differences but CTT scores did.

Figure 3.9 shows the relationship between total number correct scores and theta scores obtained from the unidimensional 3-PL model for the ELPT data. There is a positive relationship between total scores and theta scores such that they are strongly correlated, $r=.95$.

However, the scatter of the theta scores below the mean zero might imply guessing behavior since individuals with different numbers of total scores have the same theta at the lower end of the ability continuum. In addition, some examinees are omitting answers to the items, i.e. they do not answer some questions, and so the ELPT has missing data shown in Table 3.2. On the other hand, at the higher end of the ability continuum some individuals with high theta have remarkable differences in their total scores. The discrepancy between total and theta scores is further investigated in the following chapter.

CHAPTER 4

Total Scores and Theta Scores Discrepancy: Omitting Response Behavior

This chapter addresses a complication with the total number correct score approach and omitted responses given the best fitting unidimensional 3-PL model. The chapter identifies ways to investigate this complication through linear regression models. Lastly, it describes the test-taker characteristics in relationship to their omitting behavior. When the test score is the number of correct answers, it is the test-takers advantage to answer every item even if the response is at random (Lord, 1980). In the context of the dichotomous item response theory, if test-takers are instructed to use the best and optimal test taking strategy, no omitted or not-reached responses should occur. Test directions or time constraints are the factors affecting the response behavior of the test-takers. For example, responding to all unread items at random since time is running short is the best strategy in test-takers own interest. In educational tests, individuals tend to omit items when they think their response would be incorrect rather than correct. However, to account for only the targeted latent ability, respondents at a given ability level must all follow the same strategy. Thus, test instructions should be clear to explain how to act in test-takers best interest.

As already shown in Figure 3.9, which depicts the strong correlation between total and theta scores, there are some test-takers with the same theta score and dramatically different total scores. For example, some test-takers with theta score 2.2 has lower total number-correct scores than the majority with the same ability level. Table 4.1 shows numbers of correct, incorrect, and missing answers of 85 test-takers with theta of 2.2. Some test-takers have more omitted items than their ability-level peers, which yields a disadvantaged total number-correct scores for these test-takers as their ability level might be underestimated. This discrepancy between total and theta scores suggests an investigation into the missing responses in the data since ignoring omitted responses to multiple-choice items and treating

Table 4.1: *The number of correct, incorrect, and missing answers when* $\theta = 2.2$

Number of examinees	Number correct	Number incorrect	Number missing
81	69-73	7-11	0
1	69	10	1
1	69	9	2
1	70	7	3
1	63	6	11

them as wrong might yield underestimation of the test-taker ability.

The discrepancy between total and theta scores in the ELPT data suggests a connection with test-taker omitting behavior. Table 6 presents results from a linear regression (Model 1) in Equation 4.1 predicting total score from theta. Theta scores explained 89% of the variance in the total scores (R^2=0.89, $F(1,288001)$=2.38* 10^6, p=0). The theta scores significantly predicted the total number-correct scores (β_1 =14.7, p=0) . Total scores increase 14.7 points per a unit increase in the theta which is on the ability scale of the interval [-2.04, 3.44].

$$Total\,Score = \beta_0 + \beta_1\,Theta + e \tag{4.1}$$

Further investigation indicated that the residuals from the regression model of total scores on theta (Model 1) significantly correlated with the number of missing responses (r= -0.43). The plot in Figure 4.1 also shows this significant relationship between total scores and the number of missing answers. This suggested the addition of the number of missing answers into the linear regression model as it has a potential to explain more variance in total score.

Adding the number of missing responses as a predictor into the regression model to predict total scores yields Model 2 in Table 4.2.

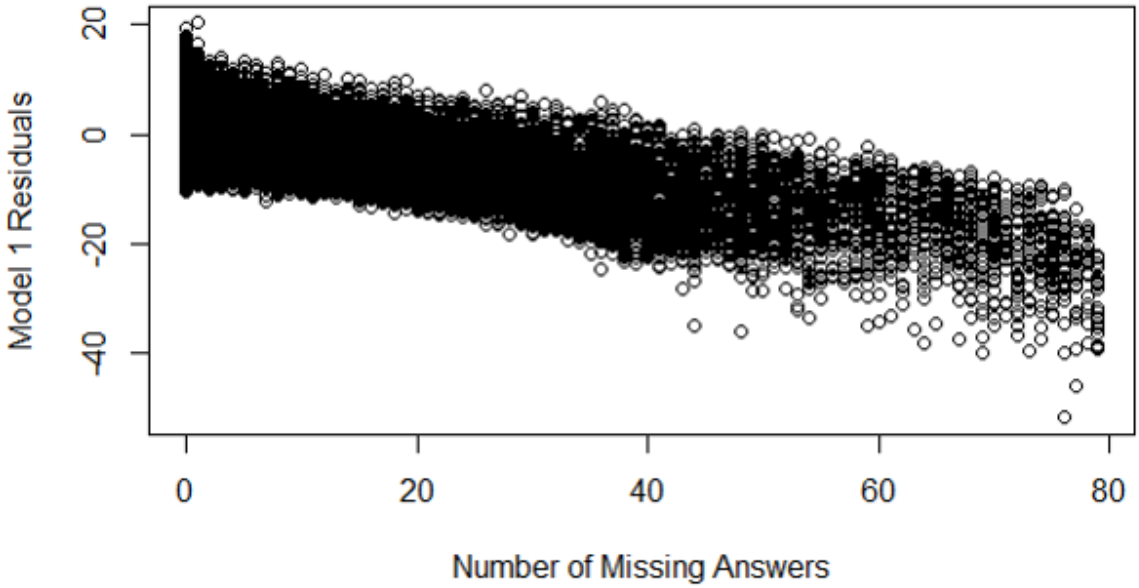


Figure 4.1: Model 1 residuals vs number of missing answers

Table 4.2: *Summary of linear regression models*

	Model 1	Model 2
Variables	Coeffient(*SE*)	Coeffient(*SE*)
Constant	30.46*(0.01)	31* (0.01)
Theta	14.72*(0.01)	14.81*(0.01)
Number of Missing Answers		-0.38 (0.00)
Theta x Number of Missing Answers		-0.18*(0.00)
R-square	0.89	0.92

$*p < .001$

$$Total\,Score = \beta_0 + \beta_1\,Theta + \beta_2\,Missing\,Answers + \beta_3\,Theta * Missing\,Answers + e \tag{4.2}$$

The model explained 92% of the variance in the total scores (R^2=0.92, F(3,287999)= 1.06*10^6, p=0). The interaction was significant as well as the main effects of the variables in Model 2. The significant interaction suggests that as the relationship between theta and total scores varies with the number of missing answers ($\beta_3 = -0.18$, p=0). For an individual with an average theta, each missing answer results in 0.38 decrease in total score

compared to other individuals with average thetas. For an individual with theta of 2, each missing answer results in 0.74 decrease in total score compared to other individuals with the same theta. Thus, the same number of missing answers at different levels of theta results in different total score. Theta moderates the relationship between the number of missing answers and total scores. As theta increases the relationship of missing answers on total scores increases. For an individual with theta of -1, each guess is worth 0.2 points but for an individual with theta of 1, each guess is worth 0.56. The higher the individual ability is, the more she could guess the omitted question correctly. This indicates that the total number-correct score might not be a fair scoring method to the ELPT given the test-takers did not equally take the advantage of the same strategy of randomly guessing on the items even though they had the same opportunity to guess. Examinees with higher ability are limiting their score potential more because they have a higher expected probability of success on each item.

Since omitting behavior yields differences in total number correct score, we examine if there are any text-taker characteristics that would correlate with missing answers. The number of missing answers did not significantly correlate with age (r=-0.04). Women (M=1.86, SD=7.10) left significantly more missing answers than men (M=1.30, SD=6.47), t(279440)= -22.29, p=0. The total score difference between men and women is also significant such that the total score of men is 0.53 higher than the total score of women. This implies that gender differences in guessing might be a good explanation for the differences in total scores.

The distribution of the number of missing answers is presented with frequencies in Table 4.3. Approximately 16.4% of the examinees omit at least one item and are likely to have a biased ranking with respect their total scores.

Another factor that might affect the total scores, theta scores, and omitting behavior is the clusters of the examinees in test centers. Some test centers might use different test instructions that would encourage to take the advantage of guessing even though the ad-

Table 4.3: *Frequency of test-takers omitting answers*

Number of missing answers	Frequency
0	240757
1-10	33530
11-20	5724
21-30	3405
31-40	1933
41-50	1119
51-60	671
61-70	476
71-80	388

ministration is simultaneous for all examinees in test centers and administration procedures are intended to be the same. Thus, the next section investigates whether the clusters yield significant correlations in total scores, theta scores, and omitted response.

4.1 Multilevel Modeling of Total Scores, Theta Scores, and Missing Answers

The clustering test-takers at the region level requires multilevel modeling to analyze the clustering of total scores, ability estimates, and omitting behavior. Since the ELPT is a large-scale paper-pencil test administered to all examinees at the same time, there might be differences in the test administrations (e.g. test instructions) at the region level. The differences might result in variation between the test centers. Thus, the purpose here is to specify correlations among responses from the same clusters and so to determine the extent of grouping effects on total scores, estimated ability, or the number of missing answers.

Clustering of total scores, theta scores, and number of missing answers is at the regional level that is either the city or the regional partitions in large cities. In the multilevel analysis, the lme4 package in R is used for the unconditional model

$$Y_{ij} = \gamma_{00} + \upsilon_{0j} + e_{ij} \tag{4.3}$$

where γ_{00} is the grand mean, υ_{0j} is the deviation between the cluster means and the

Table 4.4: *Summary of linear regression models*

	Model 1		Model 2		Model 3	
Fixed Effects	*Estimate*	*St.Err*	*Estimate*	*St.Err*	*Estimate*	*St.Err*
Intercept	28.80*	0.38	-0.11*	0.03	1.61*	0.04
	Model 1		Model 2		Model 3	
Variance Components	*Estimate*	*SD*	*Estimate*	*SD*	*Estimate*	*SD*
Intercept	7.65	2.77	0.04	0.20	0.11	0.33
Residual	204.56	14.30	0.84	0.91	45.96	6.78

$*p < .001$

grand mean, and e_{ij} is the residual individual differences from the mean of group j. The model estimates the variation in the intercept and provides the proportion of cluster level variance. Table 4.4 shows the fixed and random effects for intercept-only models at regional test centers.

The average total score is 28.80 in an interval of [0, 80] in any region and the mean of total score differs across regions within a range from 23.26 (28.80-2*2.77) to 34.24 (28.80+2*2.77) at the 0.05 level. Unexplained variation from region to region in mean total score is 7.65. The intraclass correlation coefficient for Model 1 is ρ=0.04, so 4% of the total variation in total scores is accounted for by which region each test-taker is in.

The average theta score is -0.11 in an interval of [-2.04, 3.44] in any region and the mean of theta score differs across regions within a range from -0.51 (-0.11-2*.20) to 0.29 (-0.11+2*.20) at the 0.05 level. Unexplained variation from regional center to regional center in mean theta score is 0.04. The intraclass correlation coefficient for Model 2 is ρ=0.05, so 5% of the total variation in theta scores is accounted for by which region each test-taker is in.

Correlations among responses from the same regional testing centers in their number of missing answers may have further implications about the effect of administrative or instructive differences in these centers on the test-taker omitting behavior. However, it did not appear that the number of missing answers changes significantly from center to center,

which does not indicate any significant difference in the testing patterns in regional testing centers. The average number of missing answers is 1.61 in an interval of [0, 80] in any region and the mean of missing answers differs across regions within a range from 0.95 (1.61-2*.33) to 2.27 (1.61+2*.33) at the 0.05 level. Unexplained variation from region to region in means missing answers is 0.11. The intraclass correlation coefficient for Model 3 is ρ=0, so no variation in missing answers is accounted for by which region each test-taker is in.

Therefore, total score correlations of the examinees is slightly smaller than theta correlations of the examinees within the same regional test centers, so the proportions of the variation in total scores and theta scores explained by the regional test centers are quite similar. Clustering at the regional level did not result in a significant difference in the variation of total and theta scores. Additionally, the proportion of the variance in the number of missing answers explained by the test centers is zero. Clustering at the regional level, thus, is negligible in its effect on the number of missing answers. This proves that the differences from testing center to testing center affect the total scores and theta scores at a similar level but do not affect the number of missing items. Next, the discussion of the results follows along with implications and limitations of the study.

CHAPTER 5

Discussion

This chapter discusses several conclusions about ELPT scoring within the IRT framework. Several suggestions for the ELPT developers and users along with the limitations of this study are also addressed.

CTT has been the basis for developing psychological scales and test scoring for many decades. The IRT models serve similar purposes and provide information beyond the CTT framework such that it complements CTT-based measurement approach. After fitting a unidimensional 3-PL IRT model, this study points out several conceptual and practical advantages of IRT models over CTT models in regard to the psychometric approaches to the language testing, particularly in the context of the ELPT scoring.

First, the invariant property of IRT (Hambleton et al., 1991) allows us to directly compare the scores of test-takers from the population who answer different sets of items. Assessing individuals abilities by using different test forms should not yield a difference between their ability levels given the ELPT aims to measure language ability, so the IRT model estimates the ability levels independent of different forms of the test. Given that the ELPT has examinees with missing answers, it still provides an ability estimation with comparable individual scores. The ELPT, with the current scoring method, normatively ranks the test-taker ability levels relative to total score. Total number correct scoring also ranks people on a scale from 0 to 80 with no regard to the difficulty level of items answered. Thus, the IRT framework uses distinctive parameters to formulate individual theta scores and item characteristics across subpopulations. That is, if a high-ability subpopulation is considered, all test items would not necessarily be easy in the IRT models as opposed to the CTT. Thus, the ELPT developers and users are advised to use IRT approach to the ELPT scoring rather than total number correct method. The IRT framework provides a

statistically meaningful scoring approach to compare ability as reflective of the domains of language knowledge and skills. Additionally, the test might utilize computer-based technologies to effectively cover a larger intended domain of second language ability rather than ranking the test-takers normatively mostly on their reading comprehension ability.

Second, in CTT measurement precision is the same for all scores for a particular test-taker sample and longer tests tend to be more reliable than shorter ones. However, IRT defines reliability as a function of the theta scores of the measured latent construct. The measurement precision depicted by the information functions varies across the latent construct continuum and it is generalizable to the entire population. Such individual item or entire test information functions are functions of the latent construct conditional on the item parameters. The test information function is used to evaluate the properties of the test, i.e. whether the items in a test provide adequate precision across the range of the latent construct continuum. Additionally, based on an evaluation of the item information functions and rank order of the difficulty and the discrimination parameters, the number of items in this 80-item test might be reduced in the ELPT. Item-fit of the best fitting IRT model was also investigated in this study identifying any misfitting items. Thus, ELPT developers and users can have a reduced number of items in the exam with similar measurement precision, which contributes to the efficiency of the exam.

Third, missing values are difficult to handle in CTT test scoring. Scoring by ignoring the missing responses is not accurate if examinees are not using the same test-taking strategy and this might produce differences in scores based on test-taking strategies not the intended range of latent construct. In contrast, the estimation framework of IRT models handle the analysis of the items with missing data. IRT can still calibrate items and score subjects by using all the available information based on the likelihood-based methods. Thus, this study was able to address the discrepancy between the total scores and the theta scores obtained from the 3-PL model and relate it to the omitting behavior of the test-takers. As the theta increases, examinees are more penalized in their total score with respect to the

number of missing answers. Furthermore, women tend to omit answers more than men, which produces a gender difference in total score approach to the ELPT given that omitting behavior has significant effects on the total score. Therefore, the IRT model handling the missing answers also eliminates this gender difference due to omitting behavior.

Additionally, multilevel modeling to analyze clustering resulted in similar proportions of variation in total scores and theta but not in missing answers. However, the correlations of responses from the same test centers were lower than 0.06 for both theta scores and theta. Omitting behavior did not result in a significant correlation between test-takers from the same regional test centers. That is, multilevel modeling of the number of missing answers indicated that differences in regional test centers do not significantly affect missing answers. This might suggest that there were no significant differences in test administration and test instructions from region to region.

The sample size in this study was large enough to demonstrate statistical differences between females and males based on statistical tests which will almost always result in significant differences. Even though there were statistically significant gender differences in total scores and the number of missing answers, the differences in the means of total score and in the means of the number of missing answers between males and females were small. The effect size of the gender was very small in regard to the standardized difference in the means between female and male.

There are several limitations of this study suggesting further directions. The multimodal theta distribution might be further investigated. The testlet structures in the ELPT data were not used to model the data. In addition, the test information curve in this study peaks at a very high value, which might result from neglecting the testlets. Testlet models would account for the local dependence between testlet items. Item content and item types were not analyzed in detail in this study. Before fitting the unidimensional models, it was found that there is a potential multidimensional structure in the ELPT data such that a two-dimensional model with three parameters would fit better. This second dimension,

however, needs a comprehensive investigation such as item content analysis, item types, and examinee characteristics. A multilevel IRT modeling approach might also be considered given the ELPT has clusters at the regional level.

The ELPT data has several limitations in terms of analyzing the omitting behavior. There was no response time data to investigate whether the examinees tend to omit answers because they run out of time. Since the questions and their options are shuffled for test security reasons, it was not possible to address whether the omitted answers are not-reached responses toward the end of the 80-item test since the test-takers run out of time. In addition, the ELPT is a paper-pencil test focusing primarily on the reading ability of English as a second language test-takers as opposed to a computer-based test of reading, speaking, listening, and writing.

These limitations also suggest further implications. It might be very useful to use computer-based format for the ELPT to obtain response time variables and code missing answer patterns even though the questions are shuffled. Such a format also facilitates other language domains to be measured such as listening, speaking, and writing. In addition, to have comparable language ability scores across administrations of the ELPT in different years or seasons, the IRT framework also allows test score equating. Lastly, person-fit analysis might be further investigated to detect any aberrant response behavior in ELPT data using other person-fit statistics.

References

Abedi, J. (2004). Inclusion of students with limited english proficiency in naep: Classification and measurement issues. *National Assessment Governing Board*.

Abedi, J. (2007). *English language proficiency assessment in the nation: Current status and future practice*. University of California, Davis, School of Education Davis.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., and Williamson, D. M. (2015). Learning in models with fixed structure. In *Bayesian Networks in Educational Assessment*, pages 279–330. Springer.

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

Bachman, L. F. (1997). Generalizability theory. *Encyclopedia of language and education*, 7:255–262.

Bachman, L. F. and Clark, J. L. D. (1987). The measurement of foreign/second language proficiency. *The ANNALS of the American Academy of Political and Social Science*, 490(1):20–33.

Bachman, L. F. and Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading.

Bachman, L. F. and Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL quarterly*, 16(4):449–465.

Bachman, L. F. and Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*, volume 1. Oxford University Press.

Baker, F. B. and Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, pages 395–479.

Bock, R. D. and Mislevy, R. J. (1982). Adaptive eap estimation of ability in a microcomputer environment. *Applied psychological measurement*, 6(4):431–444.

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to toefl test variance. *Language Testing*, 16(2):217–238.

Canale, M. (1983). *On some dimensions of language proficiency*. The Ontario Institute for Studies in Education.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1):1–47.

Carroll, J. B. (1961). Fundamental considerations in testing for english language proficiency of foreign students. *Testing the English proficiency of foreign students*, 36.

Carroll, J. B. (1968). The psychology of language testing. In *Language testing symposium: A psycholinguistic approach*, pages 46–69. Oxford University Press London.

Carroll, J. B. (1983). Psychometric theory and language testing. *Issues in language testing research*, pages 80–107.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29.

Choi, I. C. (1999). Test fairness and validity of the teps.language research. 35(4):571–603.

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1):67–86.

Ebel, R. L. and Frisbie, D. A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Eignor, D. R., Golub-Smith, M., and Wingersky, M. S. (1986). Application of a new goodness-of-fit plot procedure to sat and toefl item type data. *ETS Research Report Series*, 1986(2).

Embretson, S. E. and Reise, S. P. (2000). Item response theory for psychologists.

Farhady, H. (2080). *Justification, development, and validation of functional language tests*. PhD thesis, University of California at Los Angeles.

Graham, J. G. (1987). English language proficiency and the prediction of academic success. *TESOL quarterly*, 21(3):505–521.

Halliday, M. A. K. (1973). Explorations in the functions of language.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). Fundamentals of item response theory (measurement methods for the social sciences series, vol. 2).

Janssen, R. and De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research*, 34(2):245–268.

Kang, T. and Cohen, A. S. (2007). Irt model selection methods for dichotomous items. *Applied Psychological Measurement*, 31(4):331–358.

Kaplan, R. M. and Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues*. Nelson Education.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach*, volume 2. Baptism's 91 Witnesses.

Lado, R. (1961). Language testing: The construction and use of foreign language tests. a teacher's book.

Lin, T. H. and Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22(3):249–264.

Lord, F. (1952). A theory of test scores. *Psychometric monographs*.

Lord, F. (1980). Application of item response theory to practical testing problems.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39(2):247–264.

Lynch, B. K. and McNamara, T. F. (1998). Using g-theory and many-facet rasch measurement in the development of performance assessments of the esl speaking skills of immigrants. *Language Testing*, 15(2):158–180.

McCoach, D. B., Gable, R. K., and Madura, J. P. (2013). Instrument development in the affective domain. *New York, NY: Springer. doi*, 10:978–1.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, 8(1):72.

Meijer, R. R. and Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2):107–135.

Min, S. and He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, 31(4):453–477.

Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die neueren sprachen*, 75(2):165–174.

Oller, J. W. (1983). *Issues in language testing research*. ERIC.

Orlando, M. and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1):50–64.

Orlando, M. and Thissen, D. (2003). Further investigation of the performance of s-x2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27(4):289–298.

Pride, J. B. and Holmes, J. (1972). On communicative competence. In *Sociolinguistics*, pages 269–293.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, Research Report, 98-009.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stone, C. A. and Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40(4):331–352.

Sumita, E., Sugaya, F., and Yamamoto, S. (2005). Measuring non-native speakers' proficiency of english by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 61–68. Association for Computational Linguistics.

Thompson, B. and Levitov, J. E. (1985). Using microcomputers to score and evaluate items. *Collegiate Microcomputer*, 3(2):163–168.

Van Dijk, T. A. (1980). Text and context explorations in the semantics and pragmatics of discourse.

Wainer, H. and Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, 12(4):339–368.

Wainer, H. and Wang, X. (2001). Using a new statistical model for testlets to score toefl. *ETS Research Report Series*, 2001(1).

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5(2):245–262.