2012-03-16

A Comparative Analysis of Two Forms of Gyeonggi English Communicative Ability Test Based on Classical Test Theory and Item Response Theory

Young-Beol Yoon
Brigham Young University - Provo

A Comparative Analysis of Two Forms of Gyeonggi English Communicative Ability Test Based on Classical Test Theory and Item Response Theory

Young-Beol Yoon

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Richard R Sudweeks, Chair
Randall Davies
Andrew S. Gibbons

Department of Instructional Psychology and Technology

Brigham Young University

April 2012

ABSTRACT

A Comparative Analysis of Test Items of Gyeonggi English Communicative Ability Test based on Classical Test Theory and Item Response Theory

YOUNG-BEOL YOON

Department of Instructional Psychology and Technology, BYU

Master of Science

This study is an empirical analysis of the 2009 and 2010 forms of the *Gyeonggi English Communicative Ability Test* (GECAT) based on the responses of 2,307 students to the 2009 GECAT and 2,907 students to the 2010 GECAT. The GECAT is an English proficiency examination sponsored by the Gyeonggi Provincial Office of Education (GOE) in South Korea. This multiple-choice test has been administered annually at the end of each school year to high school students since 2004 as a measure of the students' ability to communicate in English. From 2004 until 2009, the test included 80 multiple-choice items, but in 2010, the length of the test was decreased to include only 50 items.

The purpose of this study was to compare the psychometric properties of the 80-item 2009 form of the test with the psychometric properties of the shorter 50-item test using both Classical Test Theory item analysis statistics and parameter estimates obtained from 3-PL Item Response Theory.

Cronbach's alpha coefficient for both forms was estimated to be .92 indicating that the overall reliability of the scores obtained from the two different test forms was essentially equivalent. For most of the six linguistic subdomains, the average classical item difficulty indexes were very similar across the two forms. The average of the classical item discrimination indexes were also quite similar for the 2009 80-item test and the 50-item 2010 test. However, 13 of the 2009 items and 3 of the 2010 had point biserial correlations with either negative or lower than acceptable positive values. A distracter analysis was conducted for each of these items with less than acceptable discriminating power as a basis to revise them.

Total information functions of 6 subdomain tests (speaking, listening, reading, writing, vocabulary and grammar) showed that most of the test information functions of the 2009 GECAT were peaked at the ability level of around $0.9 < \theta < 1.5$, while those of the 2010 GECAT were peaked at the ability level of around $0.0 \ \theta < 0.6$. Recommendations for improving the GECAT and conducting future research are included.

Keywords: CTT, IRT, test information functions, distracter analysis, English language instruction evaluation

ACKNOWLEDGMENTS

I extend a heartfelt thanks to my thesis advisor and the chair of my committee, Dr. Richard Sudweeks, for his inexhaustible expertise and inestimable advice. Special thanks to Dr. Andrew Gibbons and Dr. Randall Davies, members of my committee, for kind and encouraging comments on my thesis.

I would also like to thank to the Instructional Psychology and Technology Department faculty who shared their knowledge and wisdom so generously and helped me lead a comfortable life in a foreign country.

I would like to extend my appreciation to Dr. Sooyoung Choi and Dr. Buddy Richards who helped me start a study abroad, life-long dream, in the graduate school at Brigham Young University.

I would also like to express my gratitude to brothers and sisters of BSU and PBC in Christ; Pastor Russ Ronbinson, Dr. Eula Monroe, Dr. Shanesha, Moses Khombe and so on who kept praying for me until I finished writing this thesis.

Special thanks to Gyeonggi Provincial Office of Education. Without their scholarship program, I could never have a dream of studying abroad. As thanks, I will try to contribute to enhancing the quality of the Korean education. I'd like to also express my thanks to Seonggyu Gwon, a supervisor of GOE, who helped me collect the data of the item responses of the 2009 and 2010 GECAT for my thesis.

Above all, I wish to give my credit to my family: my wife Myeongsuk Eom, son Hyeongshik Yoon, and daughter Hyeongwon Yoon, who have led firm lives since I have been away from home.

Finally, I am really thankful to God who has made all of these completed in Him. Hallelujah! I praise the Lord with all my soul and spirit.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

As the interaction among nations increases in diverse areas, the degree of interdependence among them is also increasing. As a consequence, international cooperation is becoming more important. Due to the development of science and information technology, a move towards a knowledge and information-based society requires all the components of the society, from individual to government policies, to be able to understand and produce knowledge and information. Under these circumstances, English, being the most widely used language, is playing an important role in the communication and bonding among people with different native languages.

The dominance of English worldwide is why many non-English speaking nations emphasize English education to their elementary and secondary students. Enthusiasm for English education in Korea may be second to none in the world. To promote English education, the central government established a nationwide English curriculum that all schools are expected to implement. The objective of this curriculum is to enhance students' ability to communicate in English. In addition, the local governments strive to make English education more effective in each school. The Gyeonggi English Communicative Ability Test (GECAT) in Gyeonggi Province, with the largest number of students and teachers in Korea, is an example of the effort of the local government to improve the effectiveness of English education.

In developing a language test such as the GECAT, the first step is to specify the purpose of the test and how the scores will be used. Once the purpose of the test has been established, qualified assessment specialists are chosen to review the curriculum and establish how the concepts, knowledge, and skills will be assessed. The results are a test blueprint or

set of test specifications describing which standards and subdomains will be assessed and how the relative emphasis will be given to each of them in the test. Once this is accomplished, items are written specifically for the test. After the initial form of the test has been administered, the test results should be analyzed statistically and feedback should be provided to the developers to use as a basis for revising and improving the test. This basic developmental process should be followed for any kind of formal assessment that is proposed for widespread use.

Statement of Problem

It has been eight years since the GECAT was first developed and implemented by the Gyeonggi Provincial Office of Education (GOE) to enhance the effectiveness of English education. Every year, more than 20 teachers are mobilized to develop new test items for the GECAT. For the first six years, this test consisted of 80 items aimed at assessing students' communicative ability in (a) speaking, (b) listening, (c) reading, (d) writing, (e) vocabulary and (f) grammar. In the last two years the number of the test items was reduced from 80 to 50 at the request of the school teachers. They asked the GOE to shorten the test as the 80-item GECAT was burdensome for students to complete within the 2-hour block of time without an intermission break.

Even though the GECAT has been administered by the GOE for eight years, the test items have never been analyzed statistically. In addition, the 2010 decrease in number of test items from 80 to 50 was implemented without any analysis of the psychometric consequences of this reduction in test length. The results produced from the GECAT include the total raw scores of individual students, the percentile scores of the students within each school, and the school mean and standard deviation. These statistics are based upon the total raw score for

each student.

Classical Test Theory (CTT) has played an important role in developing and implementing tests in schools and is still widely used. However, classical test theory has some important limitations. One significant limitation is that the reported statistics are sample dependent and reflect the characteristics of the particular group of examinees as well as the characteristics of the test items.

On the other hand, item response theory (IRT), considered to be a modern test theory, produces sample invariant item statistics which are assumed to be sample independent and thus less susceptible to sampling bias which enables researchers to estimate the parameters of the test items according to the level of the ability. In this study, both CTT and IRT are used to analyze the performance of individual test items and to investigate the psychometric consequences of shortening the GECAT from 80 to 50 items.

Statement of Purpose

The main purpose of this study was to conduct an analysis of students' responses to the test items included in the 2009 and 2010 forms of the GECAT based on CTT and IRT. Another purpose was to conduct a distracter analysis based on the results of the classical item analysis and item theory analysis. The distracter analysis will provide a basis for making informed decisions about how to revise aberrant items and thereby contribute to improving the quality of the items of the GECAT.

Generally speaking, as the number of the items increases, so does the reliability under a premise that the quality of the test items is similar. If the quality of the test items of the 2009 and 2010 GECATs is similar, the reliability of the shortened 50-item test will be lower than that of the original 80-item test. The third purpose of the study was to compare the

reliability of the original 80-item test to the shortened 50-item test.

Research Questions

This study focuses on answering four research questions:

1. What are the characteristics of each item in the test in terms of classical item statistics and estimated IRT item parameters?
    a. How do the 2009 and 2010 GECAT compare in terms of classical item difficulty and item discrimination indices?
    b. How do the 2009 and 2010 tests compare in terms of estimated IRT item parameter statistics: item difficulty and item discrimination indices?
2. How well do the distracters function for items with low discrimination indices?
3. How does the reliability of scores from the shortened 50-item test compare with the reliability of the scores from 80-item test?
4. How does the test information function and the standard error of estimate for the 50-item test compare with the test information function and standard error of estimate for the original 80-item test? From these analyses we can determine how the test might be revised in order to provide more precise estimates of students' English proficiency at the location of the various cut-off points used to assign A, B, C, D, and F grades.

Chapter 2: Review of Literature

The purpose of this study is to analyze the test items in the 2009 and 2010 GECAT based on classical test theory and item response theory. The GECAT is concerned with measuring the English communicative ability on the basis of the Korea national English curriculum. This chapter considers classical test theory and item response theory, the structure of English communicative ability and the Korea national English curriculum.

Classical Test Theory

Classical test theory has provided a foundation for educational testing for a long time. The basic idea behind CTT is that the raw score (X) obtained by an individual examinee consists of a true component (T) and a random error (E) component: $X = T + E$.

The *true score* of a person can be defined as the mean score that the person would obtain on a test if they completed an infinite number of the testing sessions using the same test. Because it is not possible to obtain an infinite number of test scores, T is a hypothetical, yet central, aspect of CTT. The average of an examinee's raw scores (X_{bar}) obtained from parallel forms of the same test would be the best estimate of T.

It is also expected that the random errors around the true score will be normally distributed. If random error is normally distributed, the expected value of the error (i.e., the mean of the distribution of errors over an infinite number of trials) would be 0. In addition, those random errors are also assumed to be uncorrelated with each other; that is, there is no systematic pattern to why the raw scores of an examinee would fluctuate from time to time. Finally, random error is also assumed to be uncorrelated to the true score, T, in that there is no systematic relationship between a true score (T) and whether or not that person will obtain positive or negative errors. These assumptions about random error form the foundations of

CTT.

The standard deviation of the distribution of random errors around the true score is called the *standard error of measurement.* In CTT, standard error of measurement is a single value for all members of the tested group. The lower the value of this statistic, the more tightly packed around the true score the random errors will be.

The theory of true and error scores developed over multiple samplings of the same person held over to a single administration of an instrument over multiple persons. This new approach sped things up dramatically, because it is possible to collect data once (i.e., a single administration) on a sample of individuals (i.e. multiple persons). Therefore, a test can be given once to 2,000 students and the standard error of measurement secured from one test sitting can be generalized to the population. The equation for this process is as follows: VAR(X) = VAR(T) + VAR(E).

Given this, it can be shown that the variance of the observed scores VAR(X) that is due to true score variance VAR(T) provides the reliability index of the test. The equation for this process is as follows: VAR(T)/VAR(X) = R.

When the variance of true scores is high relative to the variance of the observed scores, the reliability (r) of the test will be high, whereas if the variance of true scores is low relative to the variance of the observed scores, the reliability (R) will be low. Reliability values range from 0.00 to 1.00. Rearranging the terms from the above equations, it can be shown that r = 1 – [VAR(E)/VAR(X)].

That is, the reliability is equal to '1 – the ratio of random error variance to total score variance.' Further, there are analyses that allow for an estimation of r (reliability). Calculating the observed variance of a set of scores is a straightforward process. Because R

and VAR(X) can be calculated, VAR (T) can be solved for with the following equation: VAR(T) = VAR(X) * R.

Item analysis. Item analysis is a process of statistically analyzing the responses of a representative group of examinees to the individual items in a test for the purpose of deciding which items function as intended by the test maker and which items need to be revised or discarded to improve the test (Henrysson, 1971; Livingston, 2006). In other words, item analysis is a formative evaluation process that allows the test maker to make informed decisions about which items need to be improved and what kinds of revisions need to be made. The item analysis process involves four main steps:

1. Constructing an initial version of the test in accordance with a set of specifications that define the main purpose for using the test and the types of cognitive tasks and subject-matter content domains from which items are to be sampled.
2. Administering the initial version of the test to a sample of examinees who are representative of the target group for whom the test is intended.
3. Computing summary statistics for each item which describe how typical examinees responded to the item and summarize the performance characteristics of the item.
4. Using the statistical results as a basis for identifying flawed items which may need to be revised or discarded.

Crocker and Algina (1986) describe a wide variety of different summary statistics that have been proposed for use in conducting item analyses. Two of the most widely used item analysis statistics based on classical test theory include the (a) item difficulty index, and (b) item-to-total correlation coefficient.

Item difficulty index. The proportion of individuals who answer a dichotomous item correct is denoted as *p* and is also conceptualized as representing the item's difficulty level. Since this statistic is a proportion, possible can range between 0.0 and 1.0. Items with high values of *p* are referred to as being relatively easy items, while those with low *p* values are considered to be relatively difficult items.

Item-to-total correlation. Another important aspect of a test item refers to the item's power to discriminate between more knowledgeable and less knowledgeable examinees. For dichotomous items, the Pearson point-biserial correlation coefficients can be used to describe this statistic. The underlying question addressed by each coefficient is how responsive an item is relative to the total test score.

The relationship between how individuals responded to each item are correlated with the *corrected* total score on the test. However, before the correlation is computed, the total score does not include the response to the item in question. This is an appropriate correction because total scores that have the item in question embedded within them will have a spuriously higher relationship than total scores made up of only the other items in the test.

This study used the adjusted (or corrected) item-total correlation form of the point-biserial correlation coefficient to estimate the discrimination index for each test item. The larger the value of a discrimination index is for an item, the better the item discriminates between low ability students and high ability students. Experience with a wide variety of classroom tests suggests that if the purpose of the test is to maximize our ability to discriminate between students' achievement, the indices of item discrimination for most of the items can be evaluated in the terms of the criteria specified in Table 1 (Ebel & Frisbie, 1986). These guidelines will be used in this study to evaluate the relative discriminating

power of individual test items.

Table 1

Classification of Test Items Based on Their Discriminating Power

Index of Discrimination	*Item Evaluation*
.40 and up	Very good items
.30 to .39	Reasonably good but possibly subject to improvement
.20 to .29	Marginal items, usually needing and being subject to improvement
Below .19	Poor items, to be rejected or improved by revision

Limitations and advantages. The main shortcoming of the classical item difficulty and item discrimination statistics is that they are sample dependent. This dependency reduces their utility. They are most useful when the examinee sample is known to be similar to the examinee population for whom the test is being developed. To the extent that the sample differs in some unknown way from the population, and this could easily happen in a field test, the utility of the item statistics may be reduced. The use of some *anchor items* in a field test that also appeared in an actual test administration can be used to partially resolve sampling problems but relationships are typically nonlinear, which complicates any such analyses.

Advantages of many classical test models are that they are based on relatively weak assumptions (i.e., they are easy to meet in real test data) and they are well-known and have a long track record. On the other hand, both individual scores and item statistics (i.e., item difficulty and item discrimination) are dependent on the test and the examinee sample, respectively, and these dependencies can limit the utility of the individual scores and item statistics in practical test development work which complicates any analyses of the results.

Item Response Theory

Item response theory became popular as a supplement and/or alternative to the classical approaches to item analysis during the late 1960s and early 1970s (Baker, 1977). Item response theory also serves a number of other important uses in test construction and scoring (de Ayala, 2009; Yen & Fitzpatrick, 2006).

Assumptions. Item response theory is based on two major assumptions: (a) unidimensionality, and (b) local item independence. These two assumptions are explored in this section.

Unidimensionality. It is commonly assumed that only one ability is being measured by a set of items in a test. Of course, this assumption cannot be strictly met because there are many cognitive, personality, and test-taking factors which impact on test performance, at least to some extent. These factors might include level of motivation, test anxiety, ability to work quickly, knowledge of the correct use of answer sheets, and other cognitive skills in addition to the dominant construct intended to be measured by the set of test items. What is required for this assumption to be adequately met by a set of test data is a *dominant component* or factor which influences test performance. This dominant component or factor is referred to as the ability measured by the test. This is the ability on which examinees are being measured. All other contributing factors to test performance are defined as errors.

Item response models in which a single ability is presumed sufficient to explain or account for examinee performance are referred to as unidimensional models. Those models that assume more than a single ability is necessary to account for examinee test performance are referred to as multi-dimensional models. These latter models are complex, and to date, not well-developed.

Local independence. Another assumption closely related to the assumption of unidimensionality is known as the assumption of *local independence* (Lord & Novic, 1968; Lord, 1980). This assumption requires that the probability of an examinee answering an item correctly (obtained from a one-dimensional model) is not influenced by his/her performance on other items in a test. When an examinee learns information from one test item helping him or her on other test items the assumption is violated. What the assumption means is that only the examinee's ability and the characteristics of the test item related to the dominant trait being measured by the test influence performance.

Features of IRT. IRT is a family of mathematical models used to analyze test item data and describe substantive characteristics of test items as well as to estimate the ability of the examinees. One distinctive feature of IRT is the idea of an item characteristic curve. Another distinctive feature is the test characteristic curve. These two different types of mathematical functions are explained next.

Item characteristic curves. An *item characteristic curve* is a mathematical function that relates the probability of answering an item correct to the ability measured by the set of items contained in the test. There is no concept comparable to the notion of an item characteristic curve in classical test theory. A primary distinction among different item response models is in the mathematical form of the corresponding item characteristic curves. It is up to the user to choose one of the many mathematical forms for the shape of the item characteristic curves. In doing so, an assumption about the items is being made that can be verified later by how well the chosen model explains the observed test results.

Each item characteristic curve for a particular item response model is a member of a family of curves of the same general form. The number of parameters required to describe the

item characteristic curves in the family will depend on the particular item response model. With the three parameter logistic model, statistics which correspond approximately to the notion of item difficulty and discrimination (used in classical test theory), and the probability of low-ability examinees answering an item correctly, are used. The mathematical expression for the three-parameter logistic curve is: $P_i(\theta) = c_i + (1 - c_i)\,[1 + e^{-Dai(\theta - bi)}]^{-1}$, i=1, 2, 3, - - - , n, which serves as the mathematical model linking the observable data (item performance) to the unobservable data (ability). $P_i(\theta)$ gives the probability of a correct response to item i as a function of ability (denoted θ). The symbol n is the number of items in the test. The c parameter in the model is the height of the lower asymptote of the ICC and is introduced into the model to account for the performance of low-ability examinees on multiple-choice test items. The b parameter is the point on the ability scale where an examinee has a $(1+c)/2$ probability of a correct answer. The a parameter is proportional to the slope of the ICC at the point b on the ability scale. In general, the steeper the slope, the higher the a parameter. The item parameters, b, a, and c, are correspondingly referred to as the item difficulty, item discrimination, and pseudoguessing parameters. The D in the model is simply a scaling factor. By varying the item parameters, many S-shaped curves or ICCs can be generated to fit actual test data. Simpler logistic test models can be obtained by setting $c_i = 0$ (the two-parameter model) or setting $c_i = 0$ and $a_i = 1$ (the one-parameter Rasch model). Thus, three different logistic models may be fit to the test data.

Test characteristic functions. One useful feature in IRT is the *test characteristic function*. It is the sum of the item characteristic functions that makes up a test and can be used to predict the number-correct score of examinees at given ability levels (Crocker, & Algina, 1986; Demars, 2010: Hambleton, 1989; Yen & Fitzpatrick, 2006). If the test is made up of test items that are relatively difficult, then the test characteristic function is shifted to the right

and examinees tend to have lower expected scores on the test than if easier test items are included. Thus it is possible through the test characteristic function to explain how it is that examinees with a fixed ability can perform differently on two tests measuring the same ability, apart from the ubiquitous error scores. The test characteristic function connects ability scores in IRT to true scores in CTT because an examinee's expected test score at a given ability level is by definition the examinee's true score on that set of test items.

Item information functions. In the case of the simple logistic models, item information functions show the contribution of particular items to the precision of the personal ability estimates. In general, items with high discrimination power contribute more to measurement precision than items with lower discriminating power, and items tend to make their best contribution to measurement precision around their *b* value on the ability scale.

Test information functions. Another special feature of item response models is the concept a *test information function*, denoted I(θ). It is the sum of item information functions in a test and provides estimates of the errors associated with (maximum likelihood) ability estimation, specifically,

$$SE(\theta) = [\ [I(\theta)]^{1/2}]^{-1}$$

This means that the more information provided by a test at a particular ability level, the smaller the errors associated with ability estimation. The presence of item and test information functions substantially alters the ways in which tests are constructed within an item response theory framework.

Item and test characteristic functions and item and test information functions are integral features of IRT models, and they are useful. However, the essential property of these

functions is what is important; that is, *model-parameter invariance*. It can also be shown that a person's parameters or abilities are estimated independently of the particular test items, and this is accomplished by incorporating the item statistics into the ability estimation process. Of course, the property of model parameter invariance is only obtained with models that fit the test data to which they are applied.

Like the CTT models, IRT models are in wide use in test development, equating test scores, identifying item bias, and scaling and reporting scores. These models are being used by many national and state organizations in the United States, and by large school districts.

Comparison of Classical Test Theory with Item Response Theory

Relationship between CTT and IRT. In the two previous sections, classical test theory and item response theory were explored in terms of their assumptions and features. Table 2 provides the differences between CTT and IRT (Hambleton & Jones, 1993).

The relationship between item difficulty and discrimination parameters in the CTT model and the two-parameter logistic model is discussed by Lord (1980). He shows that, under certain conditions (such as examinee performance not being affected by guessing), the item-test biserial correlation used within the framework of classical measurement theory and the item discrimination parameter of item response theory are approximately monotonically increasing functions of each other. This relationship may be represented as:

$$a_i = r_{pb} / \text{square root } (1\text{-}r_{pb}^2)$$

where a_i = item discrimination parameter value for item I used in IRT and r_{pb} = item biserial correlation. The relationship is approximate rather than accurate as a consequence of the different distributions and assigned scores of the two models. The number correct score (X) of classical test theory and the ability score (θ) of item response theory have distributions

Table 2

Main Differences Between CTT and IRT Models

Area	Classical test theory	Item response theory
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (i.e., easy to meet with test data)	Strong (i.e., more difficult to meet with test data)
Item-ability relationship	Not specified	Item characteristic functions
Ability	Test scores or estimated true scores are reported on the test-score scale (or a transformed test-score scale)	Ability scores are reported on the scale $-\infty \; to \; +\infty$ (or a transformed scale)
Invariance of item and person statistics	No-item and person parameters are sample dependent	Yes-item and person parameters are sample independent, if model fits the test data.
Item statistics	*p*, r_{pb}	*b, a,* and *c* (for the three-parameter model) plus corresponding item information functions
Sample size (for item parameter estimation)	200 to 500 (in general)	Depends on the IRT model but larger sample, i.e., over 500, in general, is needed.

with different shapes, and the relationship between X and θ is nonlinear. Furthermore, the total test score X is subject to errors of measurement, whereas the ability score θ is not.

Lord (1980) describes a similar monotonic relationship between p_i and b_i when all items are equally discriminating (such as in the Rasch model) so that as p_i increases b_i decreases. If items have unequal discrimination values, then the relationship between p_i and b_i will depend on r_i (Lord, 1980). This relationship may be represented as:

$$b_i \cong \gamma_i / r_i$$

where

b_i = item difficulty parameter value for item i used in IRT and

γ_i = normal deviate corresponding to the ability score beyond which p_i of the examination sample falls.

Perhaps the most important distinction between classical and modern test theories is that inherent within item response theory is the property of invariance of both item parameters and ability parameters. The consequences of this property are (a) those parameters that characterize an examinee are independent of the test items from which they are calibrated and (b) those parameters that characterize an item are independent of the ability distribution of the set of examinees (Hambleton, Swaminathan & Rogers, 1991).

Limitations of CTT and IRT. Classical test theory has a number of important limitations (Hambleton, Swaminathan, & Rogers, 1991). First and foremost is that the two statistics (item difficulty and item discrimination) that form the cornerstones of many CTT analyses are group dependent. Thus, the p and r_{pb} values, so essential in the application of CTT models, are entirely dependent on the examinee sample from which they are obtained.

In terms of discrimination indices, higher values will be obtained from examinee samples of above-average ability and lower values from examinee samples of below-average ability (Hambleton, 1989).

Another limitation of classical test theory is that scores obtained by CTT applications are entirely test dependent. Consequently, test difficulty directly affects the resultant test scores. This is an important shortcoming because the practical constraints of measurement practice frequently necessitate that examinees from a single population be compared using results obtained from different test items as a result of having been administered different forms of the same test, or at least different subtests. Indeed, CTT may be described as "test based," whereas IRT may be described as "item based." The true-score model upon which much of CTT is based permits no consideration of examinee responses to any specific item. Consequently, no basis exists to predict how an examinee, or a group of examinees, may perform on a particular test item. Conversely, IRT allows the measurement specialist greater flexibility. A broader range of interpretations may be made at the item level. Thus, item response theory permits the measurement specialist to determine the probability of a particular examinee correctly answering any given item. This has obvious advantages if a test developer needs to know the characteristics of the test scores of one or more examinee populations. Similarly, if it is necessary to design a test with particular inherent characteristics for a specific examinee population, item response models permit the test developer to do just that (Hambleton, Swaminathan, & Rogers, 1991). The need to build such tests is common: for example, a test built to discriminate among less able students to select candidates for limited special needs resources or a test built to discriminate among more able students for the award of a scholarship. In particular, this property of IRT is invaluable for certain modern testing applications, such as computerized adaptive testing.

Item response models have technical and practical shortcomings as well. On the technical side, IRT models tend to be complex, and model parameter estimation problems tend to arise in practice. Model fit can also be a problem – it is still not completely clear how problems of model fit should be addressed, especially problems that relate to test dimensionality. On the practical side, almost regardless of application, the technical demands tend to be more complex than the demands that arise with classical models. The one-parameter item response theory model certainly is more straightforward to apply than the other item response theory models (and the software, in general, is user-friendly). On the other hand, questions arise about the fit of the one-parameter model because of the restrictiveness of the model assumptions.

An awareness of the shortcomings of CTT and the potential benefits offered by IRT has led some measurement practitioners to opt to work within an IRT framework. The reason for this change of emphasis by the psychometric and measurement community from classical to item response models is as a consequence of the benefits obtained through the application of item response models to measurement problems. The four benefits of using IRT include: item statistics independent of the groups from which they were estimated, scores describing examinee proficiency not dependent on test difficulty, test models that provide a basis for matching test items to ability levels and test models that do not require strict parallel tests for assessing reliability.

Likewise, there are benefits for using the classical test theory model. Benefits obtainable through the application of CTT models to measurement problems include:

smaller sample sizes required for analyses (a particularly valuable advantage for field testing) simpler mathematical analyses compared to item response theory, model parameter

estimation is conceptually straightforward, and analyses do not require strict goodness-of-fit studies to ensure a good fit of model of the test data.

Communicative Language Ability

It is necessary to define communicative language ability prior to developing measures of language proficiency. In the early 1960s, Lado (1961) and Carroll (1961, 1968) proposed a framework for describing the measurement of language proficiency, which incorporated skills and components into their models. These models distinguished skills (i.e., listening, speaking, reading, and writing) from components of knowledge (i.e. grammar, vocabulary, phonology/graphology), but did not indicate how skills and knowledge were related.

In the mid 1970s, Hymes (1972), a sociolinguist, identified socio-cultural factors in speech and coined the term *communicative competence* for the first time. He referred to communicative competence as that aspect of our competence that enables us to convey and interpret messages and to negotiate meanings interpersonally within specific contexts.Bachman (1990) proposed components of communicative language ability in communicative language use, which incorporated the previous concepts of communicative competence.

Hymes' concept of communicative competence. The idea of communicative competence is originally derived from a distinction made by Chomsky's between *competence* and *performance*. By competence, Chomsky meant the shared knowledge of the ideal speaker-listener set in a completely homogeneous speech community. Such underlying knowledge enables a user of a language to produce and understand an infinite set of sentences out of a finite set of rules. This kind of transformational grammar provides an explicit account of tacit knowledge of language structure, which is usually not conscious but is

necessarily implicit. Hymes (1972) suggested that the transformational theory carries to its perfection the desire to deal in practice only with what is internal to language, yet to find in that internality that in theory is of the widest or deepest human significance.

Performance, on the other hand, is concerned with the process of applying innate underlying knowledge of speech to actual language use. This is commonly stated as encoding and decoding (Hymes, 1972). Because performance can never directly reflect competence except under ideal circumstances (i.e. when the ideal speaker-listener know and use language perfectly without making any mistakes), performance cannot be relevant to a linguistic theory for descriptive linguists. It involves too many performance variables to use as linguistic data, such as memory limitation, distractions, shifts of attention and interest, and errors. Therefore, according to Hymes, the most salient connotation of performance is that of an imperfect manifestation of an underlying system.

Hymes found Chomsky's distinction of competence and performance to be too narrow to describe language behavior as a whole. Hymes believed that Chomsky's view of competence was too idealized to describe actual language behavior, and therefore his view of performance was an incomplete reflection of competence. For Hymes, Chomsky's linguistic theory represented a "Garden of Eden" viewpoint that dismissed central questions of use in the area of performance. Hymes pointed out that Chomsky's theory did not account for socio-cultural factors or differential competence in a heterogeneous speech community. He also pointed out, using Labov's work, that linguistic competence covaries with the speaker. Labov described dual competence in reception and single competence in production in lower-class African-American children who distinguish Standard English and the variant Black English in recognition, but use only Black English for production. Hymes maintained that social life affects not only outward performance, but also inner competence itself. He argued that social

factors interfere with or restrict grammar use because the rules of use are dominant over the rules of grammar. Hymes further expanded this, claiming that rules of speech are controlling factors for the linguistic form as a whole.

Hymes concluded that linguistic theory must be able to deal with a heterogeneous speech community, differential competence and the role of socio-cultural features. He believed that we should be concerned with performance, which he defined as the actual use of language in a concrete situation, not an idealized speaker-listener situation in a completely homogeneous speech community. Hymes deemed it necessary to distinguish between two kinds of competence: linguistic competence that deals with producing and understanding grammatically correct sentences, and communicative competence that deals with producing and understanding sentences that are appropriate and acceptable to a particular situation. Thus Hymes coined the term *communicative competence* and defines it as knowledge of the rules for understanding and producing both the referential and social meaning of language.

Canale and Swain's concept of communicative competence. Canale and Swain (1980) believe that the socio-lingistic work of Hymes was important to the development of a communicative approach to language learning. Their work focused on the interaction of social context, grammar, and social meaning. However, as Hymes said that there are values of grammar that would be useless without rules of language use, Canale and Swain maintained that there are rules of language use that would be useless without rules of grammar. For example, one may have an adequate level of sociolinguistic competence in Canadian French just from having developed such a competence in Canadian English; but without some minimal level of grammatical competence in French, it is unlikely that one could communicate effectively with a monolingual speaker of Canadian English. However, without some minimal level of grammatical competence in French, it is unlikely that one

could communicate effectively with a monolingual speaker of Canadian French. They strongly believe that the study of grammatical competence is as essential to the study of communicative competence as is the study of sociolinguistic competence.

Canale and Swain (1980) pointed out that there is an overemphasis in many integrative theories on the role of communicative functions and social behavior options in the selection of grammatical forms, and a lack of emphasis on the role of factors such as grammatical complexity and transparency. They believed that at some point prior to the final selection of grammatical options, semantic options and social behavior options, grammatical forms must be screened for six criteria: grammatical complexity, transparency with respect to the communicative function of the sentence, generalizability to other communicative functions, the role of a given form in facilitating acquisition of another form, acceptability in terms of perceptual strategies, and degree of markedness in terms of social geographical dialects.

Furthermore, Canale and Swain (1980) point out that no communicative competence theorists have devoted any detailed attention to communicative strategies that speakers employ to handle breakdowns in communication. Examples of communication breakdowns include false starts, hesitations and other performance factors, avoiding grammatical forms that have not been fully mastered, addressing strangers when unsure of their social status, and keeping the communicative channel open. They consider such strategies to be important aspects of communicative competence that must be integrated with the other components.

Canale and Swain (1980) proposed their own theory of communicative competence and minimally included three main competencies: grammatical, sociolinguistic and strategic competence. In Canale and Swain's and later in Canale's (1983) definition, four different

components make up the construct of communicative competence (Brown, 2000). The first two components, *grammatical competence* and *discourse competence*, reflect the use of the linguistic system itself; the last two, *sociolinguistic competence* and *strategic competence* define the functional aspects of communication.

The first two components are grammatical competence and discourse competence. Grammatical competence is that aspect of communicative competence that encompasses knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics, and phonology (Canale & Swain, 1980). Discourse competence is the complement of grammatical competence in many ways, which is the ability to connect sentences in stretches of discourse and to form a meaningful whole out of a series of utterances.

The second two components are sociolinguist and strategic compentence. Sociolinguistic competence is the knowledge of the sociocultural rules of language and of discourse. This type of competence requires an understanding of the social context in which language is used: the roles of the participants, the information they share, and the function of the interaction. Strategic competence is the verbal and nonverbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or due to insufficient competence. That is, it is the competence underlying our ability to make repairs, to cope with imperfect knowledge, and to sustain communication through paraphrase, circumlocution, repetition, hesitation, avoidance, and guessing, as well as shifts in register and style.

Bachman's Communicative Language Ability. Bachman (1990) tried to incorporate a theoretical framework of language proficiency with the methods and technology involved in measuring it. He coined the term '*communicative language ability*' and proposed the

framework of communicative language ability include three components: *language competence, strategic competence*, and *psychophysiological mechanisms*.

Strategic competence is the term for characterizing the mental capacity for implementing the components of language competence in contextualized communicative language use. Strategic competence thus provides the means for relating language competencies to features of the context of situation in which language use takes place and to the language user's knowledge structures. Psychophysiological mechanisms refer to the neurological and psychological processes involved in the actual execution of language as a physical phenomenon (e.g., sound, light). The interactions of these components of CLA with the language use context and language user's knowledge structures are illustrated in Figure 1 (Bachman, 1990).

Language competence comprises, essentially, a set of specific knowledge components that are utilized in communication via language. Language competence is classified into *organizational competence* and *pragmatic competence*. Organizational competence comprises those abilities involved in controlling the formal structure of language for producing or recognizing grammatically correct sentences, comprehending their propositional content, and ordering them to form texts. These abilities include *grammatical competence* and *textual competence*. Grammatical competence consists of a number of relatively independent competencies such as the knowledge of vocabulary, morphology, syntax, and phonology or graphology. Textual competence includes the knowledge of the conventions for joining utterances together to form a text, which is essentially a unit of language – spoken or written – consisting of two or more utterances or sentences that are structured according to rules of cohesion and rhetorical organization (Bachman, 1990). Pragmatic competence consists of illocutionary competence, or the knowledge of the pragmatic conventions for

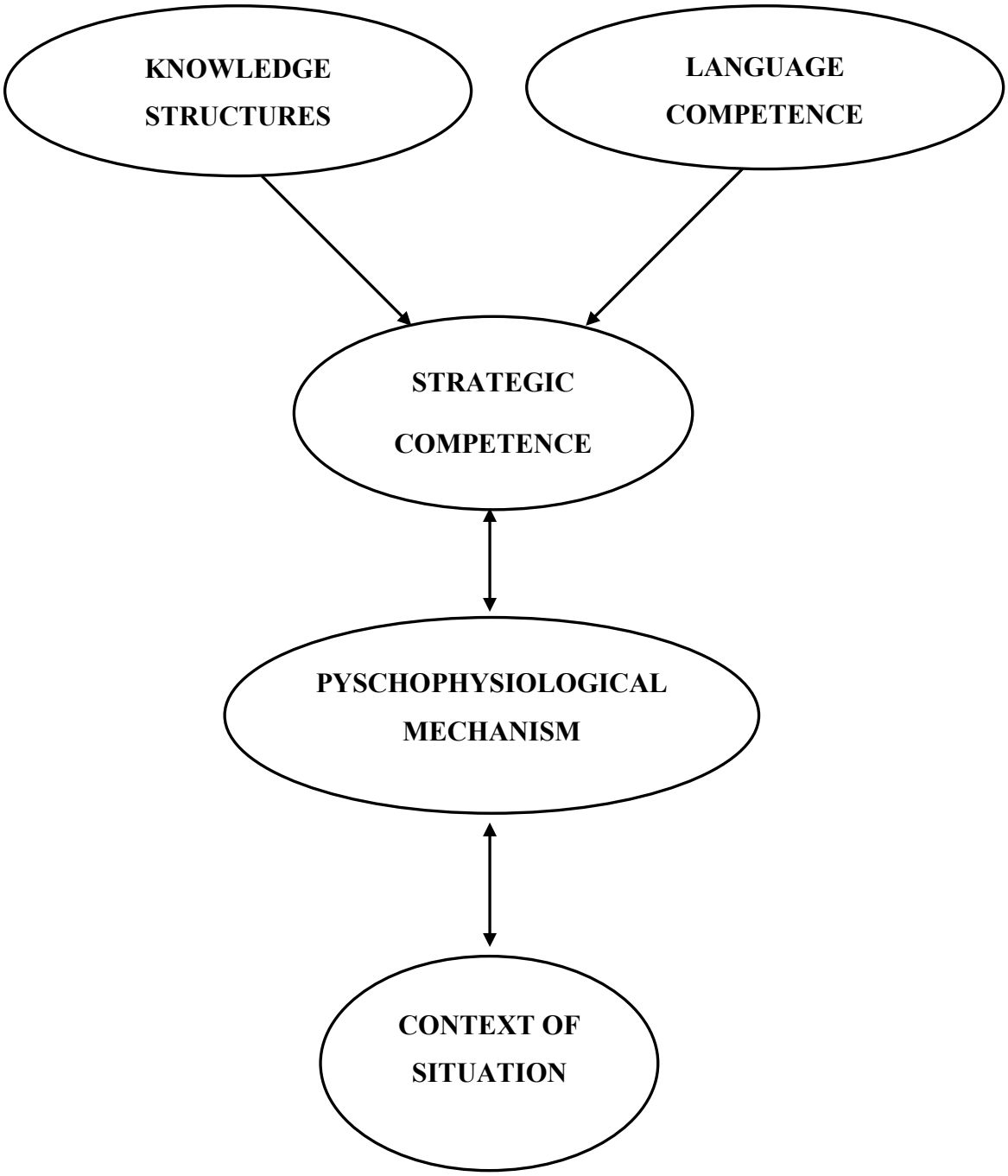


Figure 1. Components of communicative language ability in communicative language use

focus of English classes for this to be achieved.

performing acceptable language functions, and sociolinguistic competence, or knowledge of the sociolinguistic conventions for performing language functions appropriately in a give context. Pragmatics is concerned with the relationships between utterances and the acts or functions that speakers (or writers) intend to perform through these utterances, which can be called the illocutionary force of utterances, and the characteristics of the context of language use (Bachman, 1990).

The Korea National School Curriculum for English Education

The Korean government established and promulgated the first Korea national school curriculum in 1955. Since then, the curriculum has been revised nine times to be adaptable to the changes of the intra- and international surroundings. The present curriculum is known as the 2009 revised curriculum and has the individual curricula by subjects.

The curriculum is mandatory for all elementary and secondary schools. It stipulates the *characters*, the *objectives*, the *contents*, the *teaching-learning method* and *evaluation* of specific subjects. All the school activities should be based on this curriculum. Therefore, the subject of this study, the 2009 and 2010 GECAT, is also based on the 2009 revised curriculum. In this section, the Korea national English curriculum is explored, focusing on the five facets of the English curriculum for secondary schools.

Characters. The *characters* part of Korea's National School Curriculum describes the curriculum generally and broadly. First, it emphasizes the importance of learning an international language, English. The present international situation includes interdependence with the advancement of the science and information technology and active trade among countries. Communicative ability in English is important for students to survive in the present situation of the international society.

Secondly, it stipulates the characters of the elementary and the secondary English subject. It emphasizes that English, which elementary students learn from the third grade, should focus on phonetic language education. On the other hand, it stresses that secondary English education should continue to increase students' motivation, and develop the basic ability to communicate in English, while maximizing educational experiences which can increase their fluency and precision. It strongly recommends that students should become the

Lastly, the curriculum highlights the need for teaching according the different levels of the students' ability, humanity education and international culture understanding. To summarize, the *characters* portion of this curriculum guide explains the importance of English language education, its teaching and learning methods, emphasizes humanity education and international cultural understanding.

Objective. The *objective* part of the curriculum stipulates the general objective of the English education and the objectives by the level of the schools: elementary schools and secondary schools. The general objective of English education is to cultivate the basic ability to understand and use English in everyday life and to have a correct perception of foreign cultures to develop Korean culture and introduce it to other countries. Four guidelines were presented in order to achieve this. First, students should build a basis to achieve confidence to carry out life-long education in English. Secondly, students should foster the ability to communicate in everyday life and about ordinary topics. Thirdly, students should foster the ability to understand diverse foreign information and make full use of it. Finally, by understanding foreign cultures, students should newly understand Korean culture and acquire a correct perspective.

The *objective* of the curriculum stipulates the goals of elementary English and

secondary English. The objective of elementary English is to increase students' motivation to learn English and foster their basic ability to comprehend and express themselves in English. The objective of secondary English is to cultivate the ability to understand and communicate in English regarding general topics in daily life based on English learned in elementary school.

Contents. This part of the curriculum stipulates what students should be taught. The contents that students should be taught are explained with *contents system* and *accomplishment standards*.

Contents system. The contents system describes the language functions, communication activities and language materials. It emphasizes that in the sub-contents of language functions, English education gradually should foster the four language skills: speaking, reading and writing and, also build the ability to integrate the four skills. Table 3 shows the relation between language function (receptive and productive functions) and language classification (phonetic and written language).

Table 3

The relationship between language function and language classification

Language Functions	Language Classification	
	Phonetic Language	Written Language
Receptive(Comprehension)	Listening	Reading
Productive(Expression)	Speaking	Writing

The *communication activities* are explained with phonetic language and written language activities, and for the phonetic language activities, 346 exponents of seven functional categories are presented in Appendix 2 of the curriculum. These include *friendly activities*, *exchanging factual information*, *expressing cognitive attitudes*, *emotional expressions*, *expressing moral behavior*, *orders and suggestion,* and *imagining*. For the written language activities, 236 sentences of the 36 linguistic form categories are presented in Appendix 4 of the curriculum.

The *language materials* are presented with three domains: *materials*, *language*, and *vocabulary*. This part suggests materials inducing learning motivation considering the students' interest, necessity, and intellectual ability, language inducing natural language acquisition and practical communication, and the number of new words each grade may use.

Accomplishment standards. The curriculum stipulates accomplishment standards that each grade should attain. Students begin from the elementary third grade to learn English in public school and continue to the first year of high school .They are evaluated by four skills: listening, speaking, reading and writing.

The 2009 and 2010 GECAT is targeted at high schools student, so the accomplishment standards are focused on first year high school. The accomplishment standards for listening of first year high school students are based on competences. First year high school students should be able to listen to a speech or conversation on a general topic and understand the main idea and summary, listen to a speech or conversation on a general topic and understand the details, listen to and understand a short instructional broadcast, listen to a speech or conversation of differing opinions about various topics and understand the similarities and differences, listen to a simple debate and understand the main idea, and

listen to stories on various topics and understand the details of the characters.

For English speaking, first year high school students should be able to do a presentation on a familiar topic, read various stories and understand the main ideas and the summaries, exchange information about a controversial topic, read stories on various topics and express one's opinion, and change studied material into one's own words and carry out a role play. They should also be able to express themselves in various ways and carry out a simple task with instruction.

As far as English reading, first year high school students are should be able to read and understand a simple newspaper or magazine article, read a story about a general topic and understand the order of process or the logical structure, read various topics and differentiate between facts and opinions, read various topics and obtain necessary information from them, read simple stories on various topics and understand the summaries, and read various topics and understand the order of events. They should also be able to read various topics and guess what comes before and after the events, read a simple story and understand the social and cultural background and read a simple literature test and understand the main idea, characters, background, and structure.

For English writing, the first year high school students should be able to listen to a speech or conversation about a familiar, general topic and write the important information from it, read a general topic and write the summary, write simple questions, memos, and telephone messages, and write information necessary in daily routines. They should also be able to, after a trip, write a short account of it, and write about their past or future plans.

Teaching-learning method. The *teaching-learning method* of the English curriculum provides English teachers with 31 suggestions for teaching the elementary and secondary

school students. There are 14 tips for the secondary English teachers. Teachers should plan a student-centered class, where students can actively participate, and teachers can cooperate with them, develop various activities in order to achieve lively interaction between teacher and students, and among students, and use various appropriate strategies to enable students to effectively communicate. Teachers should also use audio-visual teaching materials to increase efficacy of listening activities and allow students to be naturally exposed to English phonetic language, focus on communicative activities to enhance fluency and accuracy of speaking activities and increase language ability to be applied in real circumstances, include various task-centered activities for reading, foster the ability to write the appropriate form according to the purpose. In addition, teachers should introduce various English and non-English cultures and increase the appreciation of foreign cultures and cultivate a just perspective of them. They should carry out classes in English, if possible, motivate students to get involved in learning activities with a great sense of achievement using various multime dia materials and ICTs, and operate different-leveled classes considering the circumstances of each school. Teachers should also use various methods to induce motivation and allow for a student-centered class according to students' ability, interest, and knowledge, develop various main and supplementary textbooks to accommodate individuals levels, and reorganize language functions, vocabulary, language form, etc. in developing teaching/learning materials and diversify teaching methods according the ability levels of the students.

To summarize, the teaching-learning method of the English curriculum emphasizes the student-centered class rather the teacher-centered class. It suggests the use of various teaching materials including multimedia and ICTs for inducing motivation, the operation of different-leveled classes according the ability of the students, and the cultural understanding education.

Evaluation. The evaluation of the English curriculum provides the elementary and secondary teachers with seven guidelines and nine notes in assessing students' performance and evaluating teachers' teaching process and result.

There are seven guidelines for evaluation. English teachers should establish the assessment goal according to the performance standards of the educational stage (proficiency criteria) and teaching goal before assessing, evaluate the process of teaching and the results in various methods and assess the progress of individual skills acquisition analytically or holistically. Teachers should also diagnose the students' level for applying appropriate teaching methods, implement formative assessment to check whether the methods for teaching/learning are appropriate. Teachers should use the results to improve the methods, also conduct performance assessment, if possible, in assessing productive skills such as speaking and writing, clarify the objectives, contents, types of assessment questions, and determine grading criteria before performance assessment is carried out. Teachers should also evaluate the teaching process and results through conducting evaluations of portfolios, self evaluation, and inter-peer evaluation.

The five notes for assessing the secondary school students are presented in the document. Secondary English teachers should assess the four skills: listening, reading, speaking and writing, as indicated in the curriculum. They should frequently examine the achievement of the learning objectives and analyze any reasons for depreciation in the learning process so they will not accumulate They should also assign various tasks and levels of questions in order to correctly assess students, carry out an integrated assessment in order to judge the achievement of students, and refer to this assessment to increase the effectiveness of teaching/learning methods that might enhance the students' ability to concentrate and focus on their studies.

To sum up, evaluation emphasizes not only the result assessment but also the process assessment. In addition, the establishment of the assessment goals and the clarification of the objectives, contents, types of assessment questions, and grading criteria are stressed. In particular, the evaluation of the teaching is emphasized so as to enhance the quality of teaching English. Secondary English teachers should pay attention not only to assessing the four skills using various tasks and levels of questions, but also to using the results of the assessment for enhancing the effectiveness of the teaching/learning activities.

English in Korea is learned as a foreign language and Korean English learners are not exposed to the target language as frequently as English learners of a second language. This is why the accomplishment standards are focused on the grammatical competence and the discourse competence out of Canale and Swain's (1980) communicative competence and the language competence of Bachman's (1990) communicative language ability.

Chapter 3: Method

This research is an empirical item analysis study of two versions of the GECAT that differ in length. It is designed to collect the response data of the items, to describe the statistics of the item analyses and to compare the statistics of the item analyses. Specifically, this study focused on the comparative analysis of the test items based on classical test theory and item response theory. This chapter identifies and describes the design of the study including the subjects, instruments, procedures, and analysis.

Subjects

The subject of the analysis was the test items administered in 2009 and 2010 by GOE and the number of the items was 80 and 50 respectively. The test takers' responses of the GECAT were collected with the help of the GOE. Numbers of students are on Table 4.

Table 4

Number of Students by School Year and Gender

Year	Number of schools	Number of students		
		Male	Female	Total
2009	6	1249	1108	2357
2010	7	1880	1027	2907
Total	13	3129	2135	5264

Even though the 2009 and 2010 GECATs were administered in all secondary schools, it was optional for the schools to store the students' responses to the individual test items in their computers. A GOE supervisor checked, using the telephone, for three months to

determine if the schools had the data in their scoring computers.

The six participating schools of the 2009 GECAT were (a) Guri Girls' High School, (b) Namhan High School, (c) Susung High School, (d) Suil High School, (e) Suju High School and (e) Hyoja High School. All of the students except those in Guri Girls' High School of the 2009 GECAT were freshmen. The number of the participating students of Guri Girl's High School was 475, all of whom were female sophomores.

The seven participating schools of the 2010 GECAT were (a) Gosaek High School, (b) Susung High School, (c) Suju High School, (d) Wongok High School, (e) Ichung High School, (f) Hamhyun High School and (g) Hyeonhwa High School. All of the students of the 2010 GECAT were freshmen. Susung High School and Suil High School participated in this study both in 2009 and 2010. This is why the actual number of the participant schools was 11, even though the total number of the participant schools was 13.

All the participating schools except Hyeonhwa High School were located in the urban areas, because most of the areas of the Gyeonggi Province surrounding the capital city of Korea, Seoul, were urbanized and were parts of the metropolitan area. Hyeonhwa High School was located in the urban-rural composite area.

The total number of the participant students in this study was 5264, all of whom were freshmen and sophomores from eleven high schools. The total number of the participant students of the 2009 GECAT was 2357 and that of the 2010 GECAT was 2907. The number of the male students was 3129 and that of the female students was 2135.

Instruments

The GECAT has been administered annually in all the middle and high schools of GOE since it was developed in 2004. GOE developed the test to arouse the students' interest

in learning English and to improve their English communicative ability, the objective of which was expected to correspond to that of the Korea National English curriculum.

The GECAT was purposefully designed to include test items representing six different aspects of students' proficiency in using the English language: (a) speaking ability, (b) listening ability, (c) reading ability, (d) writing ability, (e) knowledge of English grammar, and (f) knowledge of English vocabulary. The test consists solely of dichotomously scored multiple-choice items. The productive skills of writing and speaking are measured indirectly because the number of examinees is so large that it is not feasible to assess these skills directly.

The six linguistic subdomains assessed by the GECAT tests did not change when the total number of item was reduced, but the number of items within each subdomain did change. The numbers of the test items by subdomain and test year are in Table 5.

Data Analysis

Both CTT and IRT are based on the assumption that the set of test items being analyzed is *unidimensional*. In practical terms, this assumption means that the test items used to compute the examinees' total scores all measure one single trait or dimension. However, as previously explained, the items in both the 80-item 2009 GECAT and the shorter, 50-item 2010 GECAT were deliberately written to sample six different aspects of students' proficiency in using the English language.

These six different linguistic modalities include both receptive skills (reading and listening) and productive skills (writing and speaking) in addition to the two types of cognitive knowledge (vocabulary and grammar). Students' proficiency in each of these six different modalities may be positively correlated, but they are not likely to all constitute one

Table 5

Distribution of Test Items by Subdomain and Test Year

	2009 GECAT		2010 GECAT	
Subdomains	Number of Items	Percent of Total	Number of Items	Percent of Total
Speaking	18	23	7	14
Listening	14	18	15	30
Reading	28	35	17	34
Writing	4	5	5	10
Vocabulary	9	11	4	8
Grammar	7	9	2	4
Total	80	100	50	100

single trait. Theoretically, the two productive skills are more likely to be highly correlated with each other than with either of the receptive skills, and the two different types of linguistic knowledge may not be highly correlated with either of the receptive skills or with either of the productive skills. In other words, students may be highly proficient in one area without being highly proficient in another.

Consequently, the researcher assumed after careful thought that the items on both the 2009 GECAT and the 2010 GECAT were unlikely to be unidimensional and more likely to represent at least three, if not more, distinctive traits. Based on this rationale, the items in both the 2009 and 2010 versions of the GECAT were categorized into six subdomains or categories and then each category was analyzed separately rather than as a combined group. Hence, each of the six language modalities was treated as a separate subdomain. A separate analysis was then conducted for each subdomain in the 2009 test form, and a separate analysis was performed for each of the six subdomains for the 2010 test form.

The original data consisted of students' raw responses to each item in the 2009 and 2010 forms of the GECAT. Based on answer keys for each test year obtained from GOE, students' responses to each item in both forms of the GECAT were transformed into dichotomous scores ("1" if a response was correct, and "0" if the students' response was incorrect.)

The dichotomously scored data were then imported into *IBM SPSS Statistics 19* and converted into the files **.sav*. The "Reliability Analysis" routine in *IBM SPSS* was used to compute the classical item difficulty and item discrimination statistics. The "Corrected Item-Total Correlation" coefficient produced by this SPSS routine is the point biserial correlation between examinees' responses to an individual test item and their corrected total score. This correlation coefficient is labeled "corrected" because before the correlation for each item was computed, the total score for each examinee was adjusted by removing that individual's response to the corresponding item. Without this adjustment, the resulting correlation coefficient would be somewhat inflated because the uncorrected total score would include the contribution of the item with which it is being correlated. This correction procedure does not make much difference when the total scores are based on a large number of test items (e.g., 20 or more), but it is important to use corrected correlation coefficient to estimate the discriminating power of each item when total scores are computed from a small number of items (e.g., less than 20). Since many of the subdomains in the 2009 and 2010 forms of the GECAT include less than 20 items, the "Corrected-Item Total correlation" coefficient was especially relevant in the context of this study.

Since several items in both the 2009 and 2010 versions of the GECAT forms were found to have negative or low positive (< .30) corrected item-total correlation coefficients,

distracter analysis was subsequently performed on each of these aberrant items in hopes of gaining an understanding as to why these items did not function as intended and to provide a basis for making informed recommendations for improving such items and for avoiding similar problems in the process of writing similar items in the future.

To produce item parameter estimates of the subdomains in terms of IRT, first the dichotomously scored data were converted into a file with the extension *.xls into a file with the extension *.prn. Then the **.prn* file was processed in the software *BILOG-MG 3.0*. Before running *BILOG-MG 3.0*, its commands were written suitable to the purpose of the analysis, as is presented in the Appendix *M* and *N*. After saving the command file*(*.BLM*), the BILOG-*MG 3.0* was run to obtain the item parameter estimates and item information functions, and test information functions for each of the six GECAT subdomains.

After getting the original item parameter estimates, commensurable item parameters were created with a view to comparing the item parameter estimates of the subdomains of the 2009 and 2010 GECAT. For this to be achieved, the speaking subdomain items of the 2009 GECAT were used as the target metric and the others the initial metrics.

de Ayala's (2010) generic linear transformation formula was used to convert the item parameters estimates from five other subdomains to the metrics of the 2009 speaking subdomain in the commensurable metrics. The equation was $\xi^* = \zeta(\xi) + \kappa$, where ξ^* is the new parameter estimate on the target metric, ζ is the scaling constant, ξ is a parameter estimate on the initial metric that the user wants to transform, and κ is the location constant.

The approaches to computing the transformation coefficients are as follows: first, the scaling constant (ζ) is computed by dividing the standard deviation of the item location parameters on the target metric by that on the initial metric. The formulation is presented as

follows: $\zeta = S_{\delta *} / S_{\delta}$.

Second, the location constant is computed by subtracting the value of the scaling constant times the mean of the item location parameters on the initial metric from the mean of the item location parameters on the target metric. The formulation is presented as follows:

$\kappa = \delta_{bar}* - (\zeta)\ \delta_{bar}$.

Lastly, the new item discrimination parameters (α^*_j) were computed by dividing the original item discrimination parameters (α_j) by the scaling constant (ζ). The formulation is presented as follows: $\alpha^*_j = \alpha_j / \zeta$.

Chapter 4: Results

This chapter reports the results of the comparative analysis of the test items used in the 2009 and the 2010 GECAT based on classical test theory and item response theory. First, the statistical characteristics of the items in each subdomain are described in terms of classical item difficulty and item discrimination indexes and estimated IRT difficulty, discrimination, and guessing parameters. Then, the results of the distracter analysis for items with low discrimination indexes are reported. Next, the reliability of each for the six subdomains and for the composite score are described in terms of CTT. Lastly, the test information functions for each of the six domains are presented and compared across the two years.

Statistical Characteristics of the Items in Each Subdomain

This section presents the results of the analysis of the items in the 2009 and 2010 versions of the GECAT. The results for each year are presented for each of the six subdomains.

Speaking subdomain. Table 6 reports the results of the classical item analysis and the results of the IRT for each of the 18 items in the speaking subdomain in the 2009 GECAT and for each of the seven items designed to assess speaking ability in the 2010 GECAT. The average classical item difficulty statistic for the 2009 and 2010 speaking items is identical (.55). The average item difficulty parameter estimated by the IRT analysis was slightly easier for the 2010 form (.28) compared to .32 for the 2009 form, but the effect of this small difference is likely to be negligible in practice.

Comparison of the average discrimination index of the 2009 versus 2010 GECAT shows that the items in the 2010 form tend to be less discriminating on the average. Table 6

Table 6

Table 6: Speaking Subdomain Item Statistics by Estimation Procedure and Year

Test Item	Classical Item Statistics		IRT Parameter Estimates		
	Difficulty	Discrimination	Difficulty	Discrimination	Guessing
			2009 GECAT (n = 2353)		
1	.42	.44	0.72	2.63	.20
2	.73	.43	−0.39	2.09	.28
3	.40	.43	0.80	3.66	.21
4	.47	.47	0.57	3.48	.23
5	.43	.29	1.04	2.26	.28
6	.43	.39	0.83	2.07	.23
7	.65	.42	−0.14	1.85	.23
8	.61	.40	0.16	1.96	.28
9	.44	.33	0.89	1.57	.23
10	.29	.26	1.48	2.28	.19
11	.61	.42	0.11	1.90	.27
12	.78	.43	−0.78	2.39	.18
13	.62	.51	−0.24	2.41	.11
14	.80	.43	−0.88	2.55	.17
15	.50	.40	0.63	2.60	.29
16	.58	.37	0.30	1.51	.27
17	.44	.40	0.73	2.01	.21
18	.73	.38	−0.70	1.44	.16
Average	.55	.40	0.32	2.26	.22
			2010 GECAT (n = 2905)		
1	.53	.06	0.92	0.22	.13
2	.65	.44	−0.29	2.30	.11
3	.67	.33	−0.26	1.49	.19
4	.58	.38	0.23	1.83	.22
5	.27	.28	1.38	3.11	.14
6	.49	.38	0.46	1.94	.17
7	.68	.40	−0.45	1.73	.11
Average	.55	.32	0.28	1.80	.15

also reports the IRT guessing parameter for 2009 and 2010 versions of the GECAT. On the average, the seven items in the 2010 test were slightly easier, less discriminating, and less susceptible to guessing than the 18 items in the 2009 GECAT.

In general, the classical item statistics and the IRT parameter estimates both indicated that the items in both GECAT forms functioned in an acceptable manner. The only exception is Item 1 in the 2010 GECAT. Both the classical and the IRT difficulty statistics for this item are well within an acceptable range and so is the estimated guessing parameter. However, both the classical discrimination index and the estimated IRT discrimination parameter are less than acceptable and provide evidence that this item should either be revised or replaced.

Listening subdomain. The 2010 form of test for the Listening subdomain included 15 (30%) of the 50 items in the 2010 GECAT but only 14 (18%) of the items in the 2009 GECAT. The classical item statistics and the estimated item parameters for the Listening subdomain are reported in Table 7. Comparison of the average difficulty statistics for the 2009 form and the 2010 form revealed that the two forms are essentially equally difficult as indicated both by the average classical item difficulty index for the two tests and by the average IRT difficulty parameter for the two forms.

The average classical discrimination indices for the two tests are also equivalent, but a comparison of the average IRT discrimination parameter for those two forms indicated that the 2010 form had more discriminating items on the average than the 2009 form. All of the items in 2009 version and all but one of the items in the 2010 version manifested acceptable statistics.

The only exception is Item 20 in the 2010 GECAT. This item has the lowest classical discrimination index among all of the Listening subdomain items in either 2009 or 2010.

Table 7

Listening Subdomain Item Statistics by Estimation Procedure and Year

Test Item	Classical Item Statistics: Difficulty Index	Classical Item Statistics: Discrimination Index	IRT Parameter Estimates: Difficulty Index	IRT Parameter Estimates: Discrimination Index	IRT Parameter Estimates: Guessing
		2009 GECAT (n = 2355)			
19	.41	.54	0.62	1.82	.15
20	.62	.42	−0.51	1.01	.25
21	.58	.44	−0.33	1.02	.20
22	.50	.33	1.18	1.37	.35
23	.46	.50	0.34	1.41	.16
24	.48	.44	0.57	1.34	.24
25	.44	.48	0.55	1.34	.17
26	.65	.30	−1.18	0.52	.18
27	.48	.46	0.01	0.87	.10
28	.51	.42	0.44	1.23	.26
29	.44	.33	1.45	1.15	.28
30	.57	.37	0.40	1.06	.32
31	.50	.38	0.67	1.01	.27
32	.51	.44	−0.23	0.79	.10
Average	.51	.42	0.28	1.14	.22
		2010 GECAT (n = 2896)			
8	.60	.44	−0.20	1.56	.13
9	.70	.49	−0.50	2.31	.19
10	.67	.48	−0.52	1.89	.10
11	.54	.44	0.06	1.55	.14
12	.49	.36	0.21	1.08	.06
13	.54	.48	0.22	2.32	.24
14	.41	.26	1.16	1.66	.27
15	.39	.42	0.73	2.70	.18
16	.55	.57	−0.10	2.46	.11
17	.52	.36	0.39	1.56	.25
18	.56	.47	0.12	2.07	.23
19	.60	.54	−0.07	2.97	.23
20	.29	.19	2.14	0.79	.09
21	.41	.39	0.70	1.66	.16
22	.56	.41	−0.07	1.31	.09
Average	.52	.42	0.28	1.86	.16

Because of the lowest discriminating power of this item, a distracter analysis was conducted in the hopes of gaining insight into why this item did not discriminate well.

Reading subdomain. The item statistics for the Reading subdomain are reported in Table 8. Comparison of the average IRT difficulty parameter for the 2009 GECAT (0.28) and for the 2010 GECAT (0.28) indicates that the two tests are equally difficult on the average, although comparison of the average classical item difficulty indexes indicates that the 2010 version was easier.

A simple comparison of the average IRT discrimination parameter estimates for the 2009 and 2010 test forms indicated that the 2009 form is more discriminating on the average (1.44) than the average (1.37) for the 2010 form. However, this simple comparison of the mean discriminating parameter overlooks the difference in the variability of the discrimination parameters from one item to the next within each test year. The 27 estimated discrimination parameters for the 2009 form range from 1.29 to 1.74, while the estimated discrimination parameters for 17 items in the 2010 form range from 0.72 to 1.89. Hence, the variability is greater for the 17 items in 2010 than for the 28 estimated items in the 2009 test. Note that Item 18 in the 2010 test is an outlier in terms of its estimated discriminating power. This item is the only item in either form that has a discriminating power less than 1.0. However, even if this outlier is ignored, the remaining items in the shortened 2010 form are still more variable.

As shown in Table 8, the most notable problem with the items assessing the Reading subdomain is the negative value of the classical discriminating index for Item 62 on the 2009 form. Because of this negative value for the classical discrimination index, the *BILOG-MG*

Table 8

Reading Subdomain Item Statistics by Estimation Procedure and Year

Test Item	Classical Item Statistics		IRT Parameter Estimates		
	Difficulty Index	Discrimination Index	Difficulty Index	Discrimination Index	Guessing
			2009 GECAT (n = 2339)		
49	.48	.20	0.09	1.29	.14
50	.56	.48	−0.55	1.61	.12
51	.55	.43	−0.62	1.48	.06
52	.39	.22	0.87	1.35	.30
53	.65	.43	−0.85	1.59	.11
54	.34	.43	0.25	1.43	.13
55	.59	.49	−0.64	1.74	.13
56	.58	.41	−0.30	1.46	.28
57	.60	.40	−0.68	1.43	.13
58	.35	.35	0.42	1.37	.16
59	.39	.40	0.28	1.45	.22
60	.33	.37	0.49	1.45	.21
61	.55	.40	−0.48	1.41	.15
62	.17	−.08	NE[a]	NE[a]	NE[a]
63	.40	.36	0.36	1.40	.24
67	.46	.33	0.31	1.40	.31
68	.52	.34	−0.01	1.38	.29
69	.35	.36	0.47	1.46	.23
70	.45	.44	−0.12	1.45	.14
71	.44	.38	0.25	1.48	.30
72	.25	.12	1.47	1.35	.22
73	.24	.32	0.73	1.48	.15
74	.39	.33	0.49	1.44	.28
75	.25	.26	0.82	1.44	.18
76	.36	.33	0.48	1.38	.21
77	.23	.04	1.29	1.37	.21
78	.25	.16	1.06	1.40	.21
79	.32	.09	1.81	1.34	.28
Average	.41	.31	0.28	1.44	.20
			2010 GECAT (n = 2896)		
23	.75	.41	−1.18	1.34	.09
24	.63	.44	−0.67	1.20	.05
25	.61	.49	−0.48	1.42	.06
26	.54	.49	−0.24	1.26	.03
27	.54	.44	−0.12	1.12	.07
28	.54	.28	0.22	0.72	.16
29	.45	.37	0.70	1.13	.16
30	.52	.44	0.30	1.42	.18
31	.57	.32	0.01	1.08	.15
32	.48	.48	0.35	1.50	.14
33	.38	.39	0.99	1.35	.15
36	.37	.26	1.64	1.20	.23
37	.40	.50	0.66	1.73	.12
38	.46	.45	0.45	1.35	.14
39	.42	.49	0.64	1.86	.15
43	.38	.42	0.97	1.75	.17
44	.45	.48	0.58	1.89	.18
Average	.50	.42	0.28	1.37	.13

[a] Not Estimable.

software was unable to compute an estimate of the IRT discrimination parameter for this item.

A negative value for the classical discrimination index indicates that the students' response to this item was negatively correlated with their total score on all the other Reading items that year. In other words, students who were most proficient at reading English were less likely to answer this item correct than the least proficient readers. This result indicates that there may be something is wrong with this item, however, it is not unusual for an item of this difficulty to have a near zero discrimination power.

One possible reason for this undesirable result is that the answer key for this item was incorrect. Another possible reason for the negative discriminating index is that something about the wording of the item either provides a clue to the least proficient student or causes the most proficient readers to misinterpret the item. In any case, this is an undesirable result and needs to be further investigated.

When the answer key for Item 62 was checked, it was found to be correct. So a distracter analysis was conducted in hopes of obtaining more evidence about why the students' responses to this item are negatively correlated with their reading proficiency as measured by the other items in this subdomain. Distracter analyses were also conducted for Items 49, 77, and 70 on the 2009 GECAT form and for Item 28 on the 2010 form. The results of these analyses are presented later in this chapter.

Writing subdomain. The Writing subdomain was represented by only 4 (5%) of the 80 items on the 2009 GECAT, but it was represented by 5 (10%) of the items on the 2010 form. Hence, the Writing subdomain is another instance where the content representativeness of a domain increased from 2009 to 2010.

Comparison of the average classical difficulty index indicates that the 2010 form is easier, but comparison of the estimated IRT difficulty parameters provides evidence that the items on the two forms essentially equivalent in terms of average difficulty. See Table 9.

Table 9

Writing Subdomain Item Statistics by Estimation Procedure and Year of Administration

Test Item	Classical Item Statistics		IRT Parameter Estimates		
	Difficulty Index	Discrimination Index	Difficulty Index	Discrimination Index	Guessing
			2009 GECAT (n = 2349)		
64	.57	.25	−0.49	1.77	.04
65	.43	.26	−0.09	1.77	.03
66	.25	.13	0.90	1.66	.12
80	.27	.13	0.82	1.66	.12
Average	.38	.19	0.29	1.72	.08
			2010 GECAT (n = 2903)		
34	.45	.36	−0.33	1.35	.02
35	.48	.28	1.41	1.22	.07
40	.44	.20	0.34	1.18	.20
41	.42	.36	−0.20	1.34	.02
42	.38	.25	0.20	1.20	.08
Average	.43	.29	0.28	1.26	.08

The average values of the classical discrimination indices and the estimated IRT discrimination indices are inconsistent. The average IRT discriminating power is higher (1.72) for the 2009 form compared to 1.26 for the 2010 form. On the other hand, the .29 average classical discrimination index for the 2010 form is higher compared to .19 for the

2009 form. Since the IRT parameter estimates are supposedly less dependent on the particular sample of examinees who completed the test than the classical statistics, the evidence provided by the IRT parameter estimates are likely to be more credible.

Items 66 and 80 on the 2009 each have lower than acceptable classical discrimination indices, but the estimated IRT discriminating indices for both of these items were acceptable. Because of the relatively low classical discrimination value, a distracter analysis was conducted for items 64, 65, 66, and 80.

Vocabulary subdomain. The item statistics for the Vocabulary subdomain are reported in Table 10. Item 35 had a negative value for the classical item discrimination index, and Item 36 has a zero value for this statistic. The *BILOG-MG* parameter estimate algorithm did not reach convergence when these two items were included in the analysis. Consequently, Items 35 and 36 were excluded and the analysis successfully converged without them. However, all seven of the remaining Vocabulary items in 2009 GECAT have unacceptably low values of the classical discrimination index. This lack of inter-correlation provides evidence that the nine Vocabulary items in the 2009 GECAT may not be measuring the same trait. This observed heterogeneity also helps to explain why the estimated value of Cronbach's reliability coefficient that is reported later in this chapter is so low.

The four vocabulary items on the 2010 GECAT tend to have higher classical discrimination indices on the average than the 2009 items, but they are still lower than desirable and one of them (Item 45) has a negative value. Consequently, *BILOG-MG* was unable to estimate the IRT discrimination power of Item 45. *BILOG-MG* did provide estimates of the IRT discrimination parameters for the other three Vocabulary items on the 2010 form, but they are all considerably lower than desirable.

Table 10

Vocabulary Subdomain Item Statistics by Estimation Procedure and Year of Administration

Test Item	Classical Item Statistics		IRT Parameter Estimates		
	Difficulty Index	Discrimination Index	Difficulty Index	Discrimination Index	Guessing
			2009 GECAT (n = 2352)		
33	.52	.22	−0.68	2.05	.06
34	.29	.06	0.96	1.95	.24
35	.33	−.09	NE[a]	NE[a]	NE[a]
36	.15	.00	NE[a]	NE[a]	NE[a]
37	.31	.07	0.25	1.94	.15
38	.31	.16	−0.13	2.03	.05
39	.25	.15	0.23	1.97	.10
40	.30	.15	0.00	1.96	.07
41	.33	.02	1.37	1.96	.30
Average	.31	.08	0.28	1.98	.14
			2010 GECAT (n = 2902)		
45	.15	−.03	NE[a]	NE[a]	NE[a]
46	.45	.19	0.66	0.33	.08
47	.41	.26	0.70	0.33	.03
48	.49	.24	−0.51	0.33	.03
Average	.38	.17	0.28	0.33	.04

[a] Not estimable.

Because of the low observed values of the discrimination indices for the Vocabulary items, distracter analyses were conducted for all nine items on 2009 form of the GECAT and all four items on the 2010 form of the GECAT. The results are reported in this later.

Grammar subdomain. Table 11 reports the statistical characteristics of the items in the Grammar subdomain. Seven (9%) of the items on the 2009 GECAT were classified as assessing the students' knowledge of English Grammar. In contrast, only 2 (4%) of the items on the 2010 form were designed to assess knowledge of Grammar.

Table 11

Grammar Subdomain Item Statistics by Estimation Procedure and Year

Test Item	Classical Item Statistics		IRT Parameter Estimates		
	Difficulty Index	Discrimination Index	Difficulty Index	Discrimination Index	Guessing
			2009 GECAT (n = 2350)		
42	.22	.20	0.01	2.44	.03
43	.44	.08	0.04	2.35	.30
44	.24	.07	0.55	2.38	.18
45	.32	.04	0.57	2.37	.27
46	.51	.13	−0.62	2.41	.07
47	.17	−.01	1.53	2.36	.16
48	.36	.13	−0.09	2.37	.15
Average	.32	.09	0.28	2.38	.17
			2010 GECAT (n = 2901)		
49	.33	.03	NE[a]	NE[a]	NE[a]
50	.33	.03	NE[a]	NE[a]	NE[a]
Average	.33	.03			

[a] Not Estimable.

When a test that is dichotomously scored consists of only two items, there are only four possible response patterns. If a correct answer is coded as 1 and an incorrect answer is coded as 0, the only possible response patterns are 00, 10, 01 and 11. However, *BILOG-MG* is unable to estimate item parameters when the response patterns "00" and "11" are present. Consequently, no IRT parameter estimates are reported in Table 11 for the two Grammar items in the 2010 GECAT.

Most of the items in the Grammar subdomain on the 2009 and 2010 forms of the GECAT have low values for the classical discrimination index. For some reason, all of the items on the 2009 GECAT have estimated IRT parameters that exceed 2.0 in spite of the fact that all of these items have low (< .20) values on the classical discriminating index. This finding was surprising, so a distracter analysis was conducted for all seven Grammar items on

the 2009 GECAT and for both of the Grammar items on the 2010 form. The results of these distracter analyses are reported later.

The Process Used to Conduct the Distracter Analyses

Selection of marginal items. Distracter analysis was performed for items if the value of the point biserial coefficient for the item was less than .30, or if the value of the IRT discrimination parameter was less than .55. Items which fall into this category are presented in Table 12. All the items with IRT discrimination parameter less than .55 were included along with items with a point biserial correlation coefficient value less than .30. This is why the items with the *r* value less than .30 were selected as the subject items for the distracter analysis.

Table 12

Items of the 2009 and 2010 GECAT with Low Discrimination Indexes

	Items for which the point biserial correlation values are less than .30		Items for which the discrimination parameter values are less than .55	
Subtests	2009	2010	2009	2010
Speaking	-	-	.	-
Listening	-	20	-	20
Reading	49, 62, 77, 79	28	62	.
Writing	64,65,66,80	35,40,42	66	.
Vocabulary	33,34,35,36,37,38,39,40,41	45,46,47,48	34,35,36,41	45
Grammar	42,43,44,45,46,47,48	49, 50	43,47	49

Creation of ability groups for the distracter analysis. To conduct the distracter analysis, the test takers were classified into the three ability levels based on the cumulative percentile ranks. The cut-off percentile rank for classification was 33.3, but the distribution of the scores was not always like that. In consequence, the approximate value around the cut-off percentile rank was utilized. Table 13 shows the range of the percentile ranks in the Low, Middle, and High Ability Groups by sudomain and test year.

Table 13

Range of Percentile Ranks in the Low, Middle, and High Ability Groups by Subdomain and Test Year

	Low Ability Group		Middle Ability Group		High Ability Group	
	2009	2010	2009	2010	2009	2010
Speaking	0.0 - 23.5	0.0 - 25.6	23.6 - 65.0	25.7 - 60.0	65.1 - 100.0	60.1 - 100.0
Listening	0.0 - 27.4	0.0 - 33.6	27.9 - 66.1	33.7 - 64.0	66.2 - 100.0	64.1 - 100.0
Reading	0.0 - 20.6	0.0 - 52.6	20.7 - 69.5	52.7 - 80.2	69.6 - 100.0	80.3 - 100.0
Writing	0.0 - 19.4	0.0 - 37.7	19.5 - 80.1	19.5 - 79.7	80.2 - 100.0	79.8 - 100.0
Vocabulary	0.0 - 20.6	0.0 - 52.6	20.7 - 69.5	52.7 - 80.2	69.6 - 100.0	80.3 - 100.0
Grammar	0.0 - 30.1	0.0 - 45.5	30.2 - 61.1	45.1 - 88.3	61.2 - 100.0	88.4 - 100.0

Distracter Analysis for the Marginal Subdomain Items

Listening. Item 20 from the 2010 GECAT was the only item in the Listening subdomain for which distracter analysis was performed. Table 14 shows the distribution of

Table 14

Distribution of Responses to Listening Subdomain Item 20 in the 2010 GECAT

Ability Group	Option					Total
	1*	2	3	4	5	
High Ability Group	33%	9%	43%	15%	0.5%	1044
Middle Ability Group	28%	15%	40%	16%	0.3%	879
Low Ability Group	25%	19%	37%	19%	0.8%	972
Total	29%	14%	40%	17%	0.6%	2895

* Correct answer

r = .185

responses by option and ability group. In Item 20, the students were required to listen to a record speech excerpt and then answer the question "What is the best title of the talk?" The recorded audio was about 'Water Shortage Problems.' Distracter option 3 'Diseases caused by Water Pollution' was chosen more often than the correct answer. The recorded material stated water pollution as only one of the causes of water shortage. The issue of water pollution apparently was more attractive than that of water shortage to students who did not know the correct answer. Option 5 was apparently marked by mistake because the GECAT was a four-option multiple choice test.

Reading. Distracter analyses were conducted for Items 49, 62, 77, and 79 in the 2009 GECAT and Item 28 in the 2010 GECAT. The results follow.

Item 49 in the 2009 GECAT. Item 49 asked students to choose the best title of a passage written in English. The title was 'Lessening Stress Levels by Reading'. Distracter 1 was 'The Way to Read Efficiently.' Distracter 2 was 'Various Activities to Relieve Stress.'

Distracter 2 was the most attractive option to the correct answer, but Distracter 1 was not plausible except to students in the low ability group. If the result of new research conducted in Britain is not generalized to all people, Distracter 2 could be the correct answer. Item 49 could be in controversy for the correct and needs to be remedied. Distracter 4 was 'Positive Effects of Reading on Health' and was peripheral from the content even though it stated even six minutes of reading could be enough to reduce stress levels by more than two thirds.

Table 15 displays the distribution of responses to Item 49 by ability group.

Table 15

Distribution of Responses to Reading Subdomain Item 49 in the 2009 GECAT

	Option					
Ability Group	1	2	3*	4	5	Total
High Ability Group	2%	23%	66%	9%	0.0%	783
Middle Ability Group	6%	37%	44%	13%	0.0%	922
Low Ability Group	10%	37%	32%	19%	1.3%	627
Total	6%	32%	48%	13%	0.3%	2332

* Correct answer

r = .204

Item 62 in the 2009 GECAT. Table 16 shows the distribution of responses for Item 62 in the 2009 GECAT by option and ability group. Many students chose distracters. In Item 62, students were required to choose the option not related to the content of a passage written in English. The passage was about an attempt to change the diets of cows in US dairy farms to reduce the amount of methane gas. Distracter 4 was 'The amount of the greenhouse gas cows released is relatively slight.' and was chosen three times more often than the correct answer.

Table 16
Distribution of Responses to Reading Subdomain Item 62 in the 2009 GECAT

	Option					
Ability Group	1	2	3*	4	5	Total
High Ability Group	4%	3%	12%	79%	0.9%	783
Middle Ability Group	12%	13%	18%	57%	0.9%	920
Low Ability Group	16%	20%	23%	40%	1.6%	632
Total	10%	12%	17%	60%	1.1%	2335

* Correct answer

$r = -.08$

At the end of the passage it was stated that the dairy industry contributes about two percent to America's total greenhouse gas production and most of the gas comes from cows. If the students thought the amount of the greenhouse gas cows released accounted for most of the greenhouse gas the dairy industry produced, the distracter 4 could generate the

controversy of the correct answer. Consequently, the content of Distracter 4 seemed to be so vague that it was misleading to many students. Option 3, 'The American dairy is being changed into environment-friendly organic farming method', was not stated in the passage and it was the correct answer. Distracter 1 was 'US farmers are changing cows' diets to reduce the harmful gas.' Distracter 2 was 'Methane gas accelerates the global warming.' Even though the passage was written in English, all the options were written in Korean. The content of the option was translated into English by this researcher. Option 5 was chosen much more often than any other option 5 of the other items in the 2009 GECAT. If the students who chose the option 5 mistakenly had originally intended to choose Distracter 4, the proportion of the students who chose the distracter 4 would increase.

Item 77 in the 2009 GECAT. Table 17 shows the distribution of responses for Item 77 in the 2009 GECAT. Item 77 was intended to assess students' ability to draw inferences from the context presented and required students to complete a passage written in English through filling some phrases in the blank space provided in the item stem. The passage was about the use of ciphers by Samuel Pepys, a friend the king of England, Charles II. It explained that he wrote his diary in a cipher so that others could not read it.

Table 17

Distribution of Responses to Reading Subdomain Item 77 in the 2009 GECAT

	Option						
Ability Group	1	2	3	4*	5	3,4	Total
High Ability Group	13%	14%	45%	27%	0.0%	0.1%	783
Middle Ability Group	14%	25%	37%	24%	0.4%	0.0%	921
Low Ability Group	20%	27%	33%	17%	2.1%	0.0%	629
Total	16%	22%	39%	23%	0.7%	0.0%	2333

* Correct answer

r = .043

The key point of this question was 'What if Pepys' diaries have been deciphered?' and at the end of the passage was it written that 'Now that Pepys' diaries have been deciphered, they _____.' Distracter 3 'are found to be an unbreakable code even now' was so attractive as to be chosen more often than the correct answer. The correct answer was 'give us a clear picture of a period of English history.' The king of England, Charles II, was a historical character. However, it seems to be a big transition that his friend's diary written in a cipher could give us a clear picture of a period of English history with this small passage.

Item 79 in the 2009 GECAT. Table 18 shows the distribution of responses for Item 79 in the 2009 GECAT by option and ability group. Item 79, like Item 77, was a test item designed to assess the inference ability and required students to complete a passage written in English by filling in some blank phrases.

Table 18

Distribution of Responses to Reading Subdomain Item 79 in the 2009 GECAT

	Option							
Ability Group	1	2	3*	4	5	2,3	3,4	Total
High Ability Group	7%	36%	39%	17%	0.1%	0.0%	0.0%	783
Middle Ability Group	18%	25%	33%	24%	0.4%	0.0%	0.0%	920
Low Ability Group	20%	27%	23%	27%	2.4%	0.2%	0.2%	629
Total	15%	29%	32%	22%	0.9%	0.0%	0.0%	2332

* Correct answer

r = .093

The passage for Item 79 is about how to use an encyclopedia. The proportion of the High Ability Group students who chose Distracter 2 was similar to that of those who answered correctly. Distracter 2 was very attractive and the rest of the distracters functioned as they should. The words 'the first letter' of Distracter 2 might make them misleading. In the item responses of Item 79, some students chose the multiple answers: Options 2 and 3, 3 and 4, for some reason.

Item 28 in the 2010 GECAT. Table 19 shows the distribution of responses for Item 28 in the 2010 GECAT by option and ability group. Item 28 was an item that requires students to choose the main idea of the passage written in English.

Table 19

Distribution of Responses to Reading Subdomain Item 28 in the 2010 GECAT

Ability Group	Option					Total
	1	2	3*	4	5	
High Ability Group	4%	18%	74%	3%	0.0%	973
Middle Ability Group	13%	21%	54%	12%	0.1%	1214
Low Ability Group	23%	23%	27%	26%	0.8%	707
Total	12%	21%	54%	13%	0.2%	2894

* Correct answer

r = .277

The passage was about the advantage of the Internet in the newspaper company. Distracter 2 was very attractive, but Distracters 1 and 4 were not plausible for the High Ability Group students. The distribution of the responses showed Item 28 was appropriate to the Middle Ability Group students but it was easy for the High Ability Group students and difficult for the Low Ability Group students.

Vocabulary. Distracter analyses were conducted for Items 33, 34, 35, 36, 37, 38, 39, 40 and 41 in the 2009 GECAT and for Item 45, 46, 47 and 48 in the 2010 GECAT. Item 33, 34, 35, 36 and 37 in the 2009 GECAT and Item 48 in the 2010 GECAT were test items requiring students to choose an appropriate word in a blank of a short conversation, while 38, 39, 40 and 41 in the 2009 GECAT and Item 45 and 46 in the 2010 GECAT were ones requiring students to choose an appropriate word in a blank of a short sentence. Item 47 in the 2010 GECAT was a test item to requiring students to choose an appropriate word in a passage.

Item 33 in the 2009 GECAT. Table 20 shows the distribution of responses for Item 33 in the 2009 GECAT. Item 33 was a test item requiring students to choose an appropriate word in a blank of a short conversation. The student who didn't understand the context of the short conversation or the meaning of the given word chose Distracter 3 more often than any other distracter. All the distracters were not plausible for the High Ability Group students.

Table 20

Distribution of Responses to Vocabulary Subdomain Item 33 in the 2009 GECAT

	Option					
Ability Group	1*	2	3	4	5	Total
High Ability Group	82%	5%	9%	4%	0.0%	717
Middle Ability Group	50%	16%	23%	10%	0.3%	1150
Low Ability Group	12%	30%	40%	18%	0.4%	484
Total	52%	16%	22%	10%	0.2%	2351

* Correct answer

r = .218

Item 34 in the 2009 GECAT. Table 21 shows the distribution of responses for Item 34 in the 2009 GECAT by option and ability group. Distracter 2 was chosen more often than that the correct answer. Distracter 1 was attractive to the High Ability Group, while Distracter 2 to the Middle and Low Ability Group. The students who didn't understand the context of the short conversation chose Distracter 2 'advice' instead of 'regards' related to the congratulation expression.

Table 21

Distribution of Responses to Vocabulary Subdomain Item 34 in the 2009 GECAT

	Option					
Ability Group	1	2	3*	4	5	Total
High Ability Group	21%	14%	50%	15%	0.1%	717
Middle Ability Group	19%	32%	24%	24%	0.4%	1148
Low Ability Group	14%	51%	9%	25%	0.4%	484
Total	19%	31%	29%	21%	0.3%	2349

* Correct answer

r = .062

Item 35 in the 2009 GECAT. Table 22 shows the distribution of responses for Item 35 in the 2009 GECAT. Distracter 1 was chosen more often than the correct answer. Item 35 intended to asked students to respond to a request that somebody could give him a hand. The last sentence of the short sentence was 'Sure. That's ____ I am here for.' Distracter 1 was 'why', Distracter 2 'how', Distracter 3 'that', and the correct answer '4'. It seems that many students didn't understand the usage of the preposition 'for.'

Table 22

Distribution of Responses to Vocabulary Subdomain Item 35 in the 2009 GECAT

Ability Group	Option					Total
	1	2	3	4*	5	
High Ability Group	44%	7%	4%	45%	0.1%	716
Middle Ability Group	38%	17%	11%	34%	0.4%	1148
Low Ability Group	39%	28%	19%	14%	0.4%	484
Total	40%	16%	10%	33%	0.3%	2348

* Correct answer

$r = -.087$

Item 36 in the 2009 GECAT. Table 23 shows the distribution of responses for Item 36 in the 2009 GECAT by option and ability group. All the distracters were chosen more often than the correct answer. In particular, Distracter 3 was attractive regardless of the ability group. Many students didn't understand the meaning of the word 'fix' used for 'prepare food'. Distracter 2 was 'fit', Distracter 3 'mend' and Distracter 4 'repair'.

Table 23

Distribution of Responses to Vocabulary Subdomain Item 36 in the 2009 GECAT

Ability Group	Option					Total
	1*	2	3	4	5	
High Ability Group	24%	26%	37%	13%	0.1%	717
Middle Ability Group	14%	23%	44%	19%	0.4%	1148
Low Ability Group	4%	22%	38%	35%	0.8%	482
Total	15%	24%	41%	21%	0.4%	2347

* Correct answer

$r = -.002$

Item 37 in the 2009 GECAT. Table 24 shows the distribution of responses for Item 37 in the 2009 GECAT by option and ability group. Distracter 3 was chosen more than the correct answer. Over half of the High Ability Group answered correctly, while the Middle and Low Ability group chose Distracter 3. Many students didn't understand the context of this short conversation. In particular, they seemed to take 'Awesome' for the negative meaning 'Awful.' If they had taken the meaning of 'Awesome' for 'great or fantastic', they would have chosen the correct answer. The distracter 1 was 'comforts', Distracter 2 'concerns', Distracter 3 'complaints' and the correct answer 'compliments'.

Table 24

Distribution of Responses to Vocabulary Subdomain Item 37 in the 2009 GECAT

Ability Group	Option					Total
	1	2	3	4*	5	
High Ability Group	11%	11%	25%	53%	0.1%	717
Middle Ability Group	16%	21%	35%	28%	0.3%	1150
Low Ability Group	20%	22%	48%	10%	0.8%	484
Total	15%	18%	35%	32%	0.3%	2351

* Correct answer

r = .074

Item 38 in the 2009 GECAT. Table 25 shows the distribution of responses for Item 38 in the 2009 GECAT. The question was "The most _____ thing about studying in a foreign country is missing one's parents." Distracter 1 was 'confusing', Distracter 2 'rewarding', Distracter 3 'rewarding' and Option 4 'frustrating' was the correct answer.

Distracter 1 was chosen more often than the correct answer. Over half of the High Ability Group answered correctly, while the Middle and Low Ability group chose Distracter 1.

Table 25

Distribution of Responses to Vocabulary Subdomain Item 38 in the 2009 GECAT

	Option					
Ability Group	1	2	3	4*	5	Total
High Ability Group	23%	9%	10%	58%	0.0%	717
Middle Ability Group	36%	19%	20%	24%	0.4%	1147
Low Ability Group	40%	24%	28%	7%	1.4%	484
Total	33%	17%	19%	31%	0.5%	2348

* Correct answer

r = .115

Item 39 in the 2009 GECAT. The distribution of responses for Item 39 in the 2009 GECAT by option and ability group are in Table 26.

Table 26

Distribution of Responses to Vocabulary Subdomain Item 39 in the 2009 GECAT

	Option					
Ability Group	1	2*	3	4	5	Total
High Ability Group	26%	48%	8%	19%	0.0%	717
Middle Ability Group	36%	19%	18%	27%	0.5%	1150
Low Ability Group	34%	5%	27%	33%	1.0%	483
Total	32%	25%	17%	26%	0.5%	2350

* Correct answer

r = .147

The stem of this item was worded as follows: "The money that you pay for services, e.g. to a school or a lawyer, is usually called a fee or fees; the money paid for a journey is a _____." Distracter 1 was 'tip', Distracter 3 'change', Distracter 4 'payment', and the correct answer was 'fare.'Distracters 1 and 4 were chosen more often than Option 2, the correct answer. Distracter 3 was not plausible for the High Ability Group compared to Distracters 1 and 4.

Item 40 in the 2009 GECAT. Table 27 shows the distribution of responses for Item 40 in the 2009 GECAT. Again, students often chose distractor answers.

Table 27

Distribution of Responses to Vocabulary Subdomain Item 40 in the 2009 GECAT

Ability Group	Option						Total
	1	2	3	4*	5	13	
High Ability Group	7%	23%	15%	56%	0.0%	0.1%	717
Middle Ability Group	14%	33%	28%	24%	0.6%	0.0%	1150
Low Ability Group	17%	38%	39%	6%	1.2%	0.0%	482
Total	12%	31%	26%	30%	0.6%	0.0%	2349

* Correct answer

r = .153

The question was "Hospital said they could not cope with the _____ in the war." Distracter 2 was 'wound' and the correct answer was 'wounded.' Distracter 2 was chosen more often than the correct answer. Many students seemed to understand the context of the

given sentence but not to know the usage of the grammatical usage of the past participle coming after the definite article 'the.' Distracter 1 was 'weird' and Distracter 3 was 'wicked.'

Item 41 in the 2009 GECAT. Table 28 shows the distribution of responses for Item 41 in the 2009 GECAT by option and ability group. The question was "A way of life that is _____ does not need a lot of money." Distracter 1 was 'essential', Distracter 2 was 'energetic', Distracter 3 was 'electronic', and the correct answer was 'economical.' Distracter 2 'energetic' collocates with 'a way of life' but Distracter 1 and Distracter 3 don't collocate with it. Distracter 1 was the most attractive of the distracters. The proportion of the students who chose Distracter 1 was similar regardless of the student ability. These distracters should be revised.

Table 28

Distribution of Responses to Vocabulary Subdomain Item 41 in the 2009 GECAT

	Option					
Ability Group	1	2	3	4*	5	Total
High Ability Group	29%	15%	7%	49%	0.3%	717
Middle Ability Group	30%	20%	18%	32%	0.3%	1148
Low Ability Group	29%	29%	31%	10%	0.8%	483
Total	29%	20%	17%	33%	0.4%	2348

* Correct answer

r = .020

Item 45 in the 2010 GECAT. The distribution of responses for Item 45 in the 2010 GECAT by option and ability group is in Table 29**.** The question was "My girl friend always _____ me at card games." Distracter 1 was 'wins', Distracter 2 'heats', Distracter 4 'throws' and the correct answer was 'beats'. The Low Ability Group chose the correct answer more often than the Middle Ability Group. Distracter 1 was so attractive that it was chosen four times more often than the correct answer. Distracters 2 and 4 were not relatively plausible compared to Distracter 1. The discrimination value was therefore negative . It seems that many students didn't understand the usage of the word 'beat'.

Table 29

Distribution of Responses to Vocabulary Subdomain Item 45 in the 2010 GECAT

Ability Group	Option					Total
	1	2	3*	4	5	
High Ability Group	69%	6%	21%	5%	0.0%	67
Middle Ability Group	74%	7%	13%	6%	0.4%	1306
Low Ability Group	62%	11%	17%	10%	0.5%	1516
Total	68%	9%	15%	8%	0.4%	2889

* Correct answer

$r = -.034$

Item 46 in the 2010 GECAT. Table 30 shows the distribution of responses for Item 46 in the 2010 GECAT by option and ability group. The question was "Tom is not free on the 21st. We'll have to find an _____ date for the meeting." The distracter 1 was 'altering', Distracter 2 'alternate', Distracter 4 'alternant' and the correct answer was 'alternative.'

Table 30

Distribution of Responses to Vocabulary Subdomain Item 46 in the 2010 GECAT

Ability Group	Option					Total
	1	2	3*	4	5	
High Ability Group	12%	22%	49%	18%	0%	68
Middle Ability Group	11%	24%	48%	16%	1%	1305
Low Ability Group	14%	25%	41%	20%	1%	1517
Total	13%	24%	45%	18%	1%	2890

*Correct answer

r = .191

Distracter 2 was attractive but the rest of the distracters functioned as they should. The item responses were similar regardless of the options and the level of the ability. The test writers seemed to try to confuse the students with the words having the similar spelling to the correct answer. Even so, the task presented to the examinee was to check if the distracters are able to collocate with the words presented in the context of the given sentence.

Item 47 in the 2010 GECAT. The distribution of responses for Item 47 in the 2010 GECAT is shown on Table 31. The task in Item 47 was to choose the inappropriate word in

the context. Distracters 2 and 3 were equally attractive. The Middle Ability Group chose the correct answer more often that the High Ability Group. Option 5 was chosen by 1.5% of the High Ability Group. If they had taken Option 5 for the correct answer 4 by mistake, the difference of the proportion between the High and the Middle Ability Group students who answered correctly would have reversed. The inconsistency of the number of the options in the testing paper and that in the answer-sheet might have brought about this result.

Table 31

Distribution of Responses to Vocabulary Subdomain Item 47 in the 2010 GECAT

Ability Group	Option					Total
	1	2	3	4*	5	
High Ability Group	10%	22%	19%	47%	1.5%	68
Middle Ability Group	13%	18%	21%	49%	0.5%	1306
Low Ability Group	16%	25%	25%	34%	0.5%	1513
Total	14%	22%	23%	41%	0.5%	2887

* Correct answer

r = .263

Item 48 in the 2010 GECAT. Table 32 shows the distribution of responses for Item 48 in the 2010 GECAT. Item 48 required students to choose an appropriate word in a blank ***in*** a short conversation. The question was related to the grammatical structure of a sentence. The question was

A: This is _____ spicy that I cannot eat any more.

B: Oh, I didn't know that. I'll bring a cup of water.

Table 32

Distribution of Responses to Vocabulary Subdomain Item 48 in the 2010 GECAT

Ability Group	Option					Total
	1*	2	3	4	5	
High Ability Group	63%	10%	7%	19%	0.0%	68
Middle Ability Group	57%	11%	9%	24%	0.2%	1305
Low Ability Group	41%	16%	12%	30%	0.4%	1514
Total	49%	13%	11%	27%	0.3%	2887

* Correct answer

r = .243

Distracter 2 was 'much', Distracter 3 'such', Distracter 4 'very' and the correct answer was 'so.' Distracter 3 was not so plausible, while Distracter 4 was very attractive. Even though the students didn't know the meaning of the context, they could choose the correct answer only if they had the grammatical knowledge about the cause-effect sentence structure 'so ~ that … cannot.' Item 48 was not a good test item in the communicative language testing method putting an emphasis on the context.

Grammar. Distracter analyses were conducted for Items 42, 43, 44, 45, 47 and 48 in the 2009 GECAT and for Items 49 and 50 in the 2010 GECAT. Item 42 in the 2009 GECAT assesses students' ability to choose a grammatically appropriate word in a blank of a given sentence. Items 43, 44, 45 in the 2009 GECAT and Item 49 in the 2010 GECAT was test items to choose the option that contains an error in a short conversation. Items 46, 47, 48 in the 2009 GECAT and Item 50 in the 2010 GECAT were test items to choose the ungrammatical or inappropriate parts in a context-embedded passage.

Item 42 in the 2009 GECAT. The distribution of responses for Item 42 in the 2009 GECAT is shown in Table 33. The question was "The doctor recommended that I _____ less salt and sugar." Distracter 2 was 'used', Distracter 3 was 'using', the Distracter 4 was 'have used' and the correct answer was 'use'. Distracters 2 and 4 were chosen more often than the correct answer. Item 42 focussed on the grammatical syntax of the infinitive verb coming after the verbs such as *demand*, *require*, *request*, *order*, *insist*, *suggest*, *propose* and *recommend*. However, this item should be revised in subsequent tests, because the given sentence was a context-reduced one. It is irrelevant to the communicative language testing method.

Table 33

Distribution of Responses to Grammar Subdomain Item 42 in the 2009 GECAT

	Option						
Ability Group	1*	2	3	4	5	3,4,5	Total
High Ability Group	43%	21%	15%	21%	0.1%	0.1%	915
Middle Ability Group	15%	30%	24%	31%	0.1%	0.0%	727
Low Ability Group	3%	32%	27%	37%	0.7%	0.0%	706
Total	22%	27%	21%	29%	0.3%	0.0%	2348

* Correct answer

r = .199

Item 43 in the 2009 GECAT. Table 34 shows the distribution of responses for Item 43 in the 2009 GECAT by option and ability group. Item 43 focused on choosing an ungrammatical or inappropriate part in a short conversation. The question was

① A: Someone stole my wallet last night.

② B: Oh, no! What happened?

③ A: Well, I guess I'd forgotten locking the locker.

④ B: That's terrible! Did you lose a lot of money?

Table 34

Distribution of Responses to Grammar Subdomain Item 43 of the 2009 GECAT

Ability Group	Option					Total
	1	2	3*	4	5	
High Ability Group	7%	12%	66%	15%	0.0%	915
Middle Ability Group	10%	22%	45%	23%	0.0%	727
Low Ability Group	17%	32%	16%	33%	1.3%	705
Total	11%	21%	44%	23%	0.4%	2347

* Correct answer

r = .078

This question was a test item about a verb *forget* taking two kinds of objects such as *to infinitive* and *gerund* with the different meanings. It seemed that the distracters functioned as they should. Distracter 4 was attractive for the Middle Ability Group and the Low Ability Group. Distracter 1 was not plausible for the High Ability Group.

Item 44 in the 2009 GECAT. Table 35 shows the distribution of responses for Item 44 in the 2009 GECAT by option and ability group. In Item 44, students were to choose an ungrammatical or inappropriate part in a short conversation. The question was

① A: My friend and I biked out into the countryside last weekend.

② B: Wow! Do you bike often?

③ A: Yeah! We are belonged to the Weekend Biking Club.

④ B: Did you have a picnic on your last trip?

Table 35

Distribution of Responses to Grammar Subdomain Item 44 in the 2009 GECAT

Ability Group	Option					Total
	1	2	3*	4	5	
High Ability Group	7%	14%	41%	38%	0.0%	913
Middle Ability Group	12%	22%	17%	49%	0.1%	727
Low Ability Group	13%	24%	9%	53%	0.7%	705
Total	10%	20%	24%	46%	0.3%	2345

* Correct answer

r = .069

This item was about a static verb *belong to,* which is not usually used in the passive voice. Distracter 4 was so attractive that it was chosen more often than the correct answer. Distracter 4 might confuse the students with an ambiguous word *picnic* which means a kind of food eaten on the excursion. Almost as many students of the High Ability Group chose Distracter 4 as often as the correct answer. Distracter 1 was not so plausible enough to attract the students regardless of the ability level.

Item 45 in the 2009 GECAT. Table 36 shows the distribution of responses for Item 45 in the 2009 GECAT by option and ability group. The question was

① A: I was surprised at how good the weather was.

② B: Yes, it was really sunny. It was surprising.

③ A: It was good to lie in the sun. It was so relaxed.

④ B: There was a lot to see, too.

Table 36

Distribution of Responses to Grammar Subdomain Item 45 in the 2009 GECAT

	Option					
Ability Group	1	2	3*	4	5	Total
High Ability Group	10%	10%	51%	29%	0.1%	915
Middle Ability Group	14%	16%	29%	41%	0.6%	726
Low Ability Group	18%	19%	12%	48%	2.1%	704
Total	14%	15%	33%	39%	0.9%	2345

* Correct answer

r = .040

This question was about a usage of adjectives with the suffixes *-ed* and *-ing* such as *surprised*/*surprising* and *relaxed*/*relaxing*. Item 45 was also a test item to choose an ungrammatical or inappropriate part in a short conversation. Distracter 4 was so attractive that it was chosen more often than the correct answer. In particular, it was attractive for the Middle Ability Group and the Low Ability Group. Distracter 4 can generate controversy as a

correct answer. It can be asserted that the infinitive *to see* should be changed into *to be seen*. In addition, if the students marked Option 5 for Distracter 4 by mistake, the actual proportion of the response to it would have increased. Distracters 1 and 2 were not so plausible for the High Ability Group.

Item 46 in the 2009 GECAT. Table 37 shows the distribution of responses for Item 46 in the 2009 GECAT by option and ability group. This question addressed subject-verb agreement in a context-embedded passage, which could make the Low Ability Group confused. Distracters 2, 3 and 4 were chosen more often than the correct answer by the Low Ability Group. All the distracters did not function as they should. In particular, they were not plausible for the High Ability Group. It seemed that the Low Ability Group did not understand what Item 46 was asking.

Table 37

Distribution of Responses to Grammar Subdomain Item 46 in the 2009 GECAT

Ability Group	Option					Total
	1*	2	3	4	5	
High Ability Group	77%	6%	8%	8%	0.1%	913
Middle Ability Group	50%	14%	18%	18%	0.4%	727
Low Ability Group	20%	23%	28%	29%	0.9%	704
Total	51%	14%	17%	17%	0.4%	2344

* Correct answer

r = .131

Item 47 in the 2009 GECAT. The distribution of responses for Item 47 in the 2009 GECAT by option and ability group is shown in Table 38. All the distracters were chosen more often than the correct answer. Distracter 4 was so attractive that it was chosen more often than the correct answer even by the High Ability Group. It seemed that Item 47 lost its focus on its grammatical point, because its options included different grammar points. For example, the correct answer 1 was related to the active and passive voice, Distracter 2 a gerund as an object, Distracter 3 the present perfect tense and Distracter 4 the present tense followed by the verb *hope* replacing the future tense. Item 47 should be revised.

Table 38

Distribution of Responses to Grammar Subdomain Item 47 in the 2009 GECAT

Ability Group	Option					Total
	1*	2	3	4	5	
High Ability Group	27%	18%	26%	29%	0.2%	914
Middle Ability Group	16%	23%	30%	30%	0.6%	726
Low Ability Group	6%	29%	33%	31%	0.6%	703
Total	17%	23%	29%	30%	0.4%	2343

* Correct answer

$r = -.014$

Item 48 in the 2009 GECAT. Table 39 shows the distribution of responses for Item 48 in the 2009 GECAT by option and ability group**.** The distracter responses worked well.

Table 39

Distribution of Responses to Grammar Subdomain Item 48 in the 2009 GECAT

Ability Group	Option					Total
	1	2	3	4*	5	
High Ability Group	8%	14%	18%	59%	0.2%	915
Middle Ability Group	16%	18%	37%	30%	0.4%	726
Low Ability Group	21%	23%	44%	11%	0.6%	704
Total	15%	18%	32%	36%	0.4%	2345

* Correct answer

r = .125

Distracter 3 was attractive and the rest of the distracters functioned as they should. Distracter 1 was related to the grammatical point of the relative pronoun *who*, Distracter 2 to the reflexive pronoun *himself*, Distracter 3 to the predeterminer *any* and the correct answer 4 also to the predeterminer *dozens of*. Item 48 asked multiple grammatical points in a single question like Item 47.

Item 49 in the 2010 GECAT. The distribution of responses for Item 49 in the 2010 GECAT is in Table 40. In Item 49 the student chose the option that contained an ungrammatical component in the context-embedded short conversation. The question was

① A: Dear, have you seen my luck red tie?

② B: I've seen it a couple of weeks ago. Why?

③ A: I wanted to wear it to my manager's meeting.

④ B: OK. Let me help you look.

Item 49 illustrated the present perfect not used with the adverb phrase meaning the distinct past tense such as *a couple of weeks ago*. Distracter 3 was chosen by many students regardless of ability level for some reason. They might think that the tense of the verb *wanted* had something wrong. Distracter 4 was also relatively attractive. They might think that the object *you* had something wrong. Distracter 4 might also generate the controversy of the correct answer. Distracter 3 was so attractive that it was chosen as often as the correct answer, while the Distracter 1 was not plausible. The proportion of the students who chose each option was similar regardless of the ability level.

Table 40

Distribution of Responses to Grammar Subdomain Item 49 in the 2010 GECAT

Ability Group	Option					Total
	1	2*	3	4	5	
High Ability Group	7%	37%	32%	24%	0.6%	337
Middle Ability Group	8%	33%	36%	22%	0.6%	1232
Low Ability Group	11%	33%	32%	23%	0.5%	1312
Total	9%	34%	34%	23%	0.6%	2881

* Correct answer

r = .034

Item 50 in the 2010 GECAT. Table 41 shows the distribution of responses for Item 50 in the 2010 GECAT by option and ability group. Many students chose distracters.

Table 41

Distribution of Responses to Grammar Subdomain Item 50 in the 2010 GECAT

Ability Group	Option					Total
	1	2	3	4*	5	
High Ability Group	9%	19%	36%	35%	0.3%	331
Middle Ability Group	11%	21%	35%	32%	0.6%	1210
Low Ability Group	12%	20%	33%	34%	0.6%	1294
Total	12%	20%	34%	33%	0.6%	2835

* Correct answer

r = .034

Item 50 presented the examinees with a series of sentences embedded within a paragraph. The task presented to the students was to decide which one of the sentences contains a grammatical error. The resulting context-dependent item set is shown below.

> ① Michael Garland has written and illustrated many books for children.
> ② He spent his childhood in New York, exploring the woods, playing sports, and drawing. ③When he drew something in school, his teacher would often show it to the class, and put it up on the bulletin board. ④This helped him to decide what he wanted to become an artist.

For some reason, students chose Distracter 3 more often than the correct answer including the High Ability Group. On the other hand, the Low Ability Group chose the correct answer more often than the Middle Ability Group. Perhaps they thought that the auxiliary verb *would* had something wrong grammatically. Item 50 included the multiple grammatical points like

Items 46, 47 and 48 in the 2009 GECAT. Distracter analysis showed that if the Grammar subdomain test items include multiple grammatical items, the distribution of the responses to the items may be aberrant.

Reliability of Scores from the 50-Item GECAT and the 80-Item GECAT

Table 42 displays the estimated reliability coefficients as well as the number of items by subtest and testing year. In general, an increase in the number of the test items will produce an increase in reliability and a decrease in the number of the test items will lead to a decrease in the reliability. If the number of the test items of the 2009 GECAT was larger than that of the 2010 GECAT, the estimated reliability of the 2009 GECAT would be expected to be greater than that of the 2010 GECAT under the premise that the quality of the test items is

Table 42

Estimated Reliability and Number of Test Items by Subdomain and Testing Year

	Year of Test Administration			
	2009		2010	
Subdomain	Number of Items	Cronbach's Alpha	Number of Items	Cronbach's Alpha
Speaking	18	.82	7	.60
Listening	14	.80	15	.81
Reading	28	.80	17	.82
Writing	4	.36	5	.52
Vocabulary	9	.23	4	.32
Grammar	7	.23	2	.07
Combined	80	.92	50	.92

same. Even though the number of the test items of the 2010 GECAT was smaller than that of the 2009 GECAT by 30 items, the overall estimated reliability (.92) of the 2010 GECAT was identical to that of the 2009 GECAT.

The estimated reliability of the Reading subdomain items of the 2010 GECAT was the highest and that of the Grammar test items of 2010 GECAT was the lowest. In spite of the fact that the number (17 items) of the Reading subdomain items of the 2010 GECAT was smaller than that (28 items) of the 2009 GECAT, the estimated reliability (.82) of the Reading subdomain items of the 2010 GECAT was higher than that (.80) of those of the 2009 GECAT. That was similar to the findings for the Vocabulary subdomain items. The estimated reliabilities of the Vocabulary and Grammar subdomains of the 2009 and 2010 GECAT are low compared to those of the other subdomains. Ironically, the reliability (.32) of scores obtained from the four Vocabulary items in the 2010 GECAT is greater than the reliability (.23) of scores obtained from the nine Vocabulary items in the 2009 GECAT. Conversely, the reliability of scores from the Grammar test diminished from .23 to .07 when the number of GECAT items was reduced from seven to two.

The small number of 2010 items devoted to assessing knowledge of Grammar and the low reliability coefficient raises the question of whether there is any value in attempting to assess students' knowledge of Grammar in this test. Consideration should be given either to expanding the Grammar section sufficiently to provide a content-valid sample of students' knowledge of English grammar and yield reliable scores, or to deleting this section entirely.

Precision of the Person Ability Estimates

One of the main purposes for giving a test is to obtain an estimate of the degree to which each individual examinee possesses or lacks the ability or trait measured by the test

(*person ability estimate)*. An advantage of using IRT compared to CTT is that IRT provides a means of describing the degree to which each individual test item contributes to the precision (or lack of precision) of the person ability estimate for the various examinees who responded to a test.

Item Information Functions. The *BILOG-MG* software produces an Item Information Function (IIF) for each item in a test which graphically displays how the relative precision of the person ability estimates varies across different levels of ability. Appendix *A* displays an IIF for each of the 80 items in the 2009 GECAT, and Appendix *B* displays an IIF for each of the 50 items in the 2010 GECAT. The vertical axis in each of these graphs is a measure of the psychometric information (precision), and the horizontal axis represents varying levels of the particular language ability or subdomain measured by the item. The curved line in each graph describes how the precision of the ability estimates varies across the varying levels of ability ranging from very low levels of ability (– 3.0) on the left end of the continuum to average levels of ability in the middle (0.0) of the continuum to high levels of ability on the right end (3.0) of the continuum.

The IIFs in each Appendix are organized by the subdomain. Inspection of these graphs in each Appendix shows that some items are more informative (i.e., provide most information) on the right side of the horizontal axis. In other words, they provide greater precision for estimating the abilities of above average students and less information for estimating the ability of below average students. Conversely, some items provide more information for estimating the ability of below average students and less information for estimating the ability of above average students.

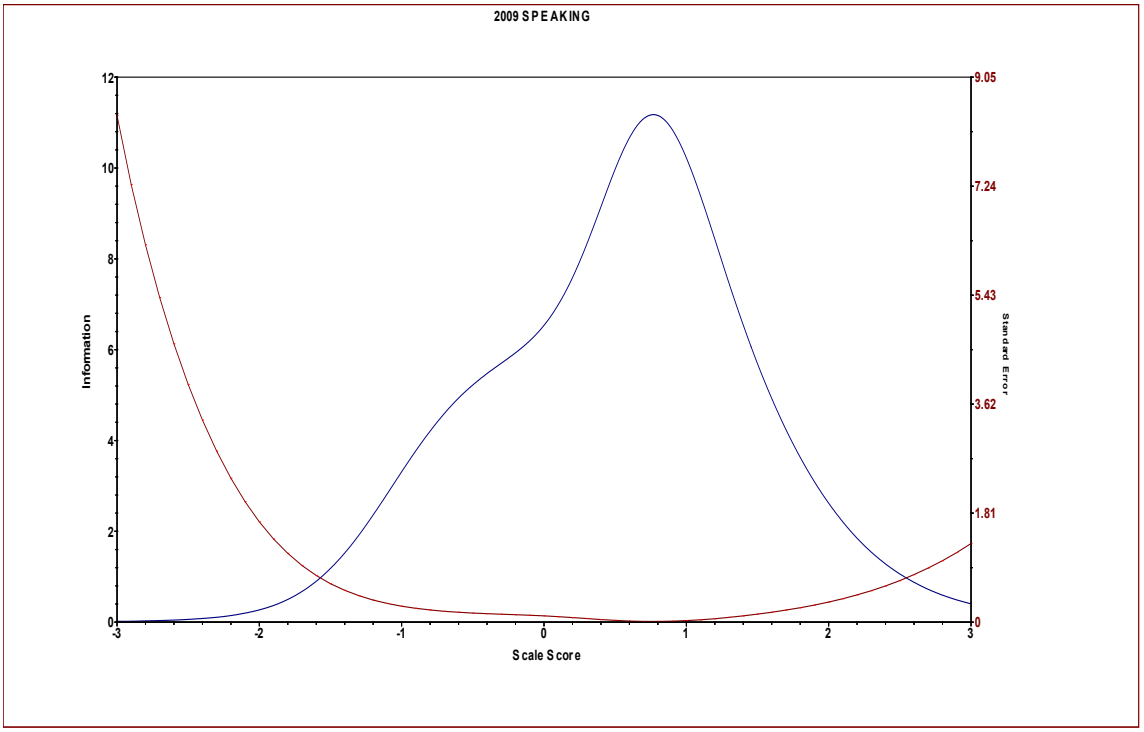
Test Information Functions. In this study, the items within each subdomain were analyzed separately. It is possible compute an estimate of each student's language proficiency

or ability on each of the six subdomains. The IIFs for the items within a particular subdomain can be summed to produce a Test Information Function (TIF) which describes how the precision of the subdomain ability estimates varies across levels of the ability continuum.

Test information for the Speaking subdomain. Figure 2 displays a TIF for the Speaking subdomain in the 2009 GECAT and another TIF for the Speaking section of the 2010 GECAT. Since the number of items in the Speaking subdomain was reduced from 18 in the 2009 GECAT to 7 in the 2010 GECAT, it is not surprising that the 2010 form produces much less information at each ability level than the 2009 GECAT. This reduction in precision is a direct result of deleting 11 items from the Speaking subdomain. If GOE administrators want to obtain more precise estimates of students' speaking proficiency they need to add at least four more items to this subdomain. In both years the items in the Speaking subdomain provide more precise estimates of ability for students in the above average range and less precise estimates for below average students. Hence, the person ability estimates will be less reliable for students whose ability to speak English is below the average of the total population of examinees who responded to the test. If the test makers want to obtain more precise estimates of ability for below average students, they need to add more items to the test that have IRT difficulty parameters that are less than 0.0.

Test information for the Listening subdomain. Figure 3 displays a TIF for the Listening subdomain in the 2009 GECAT and another TIF for the Listening in the 2010 GECAT. Even though the number of test items (80) in the 2009 GECAT was reduced to 50 in

2009 Speaking Subdomain (Number of Items = 18)



2010 Speaking Subdomain (Number of Items = 7)

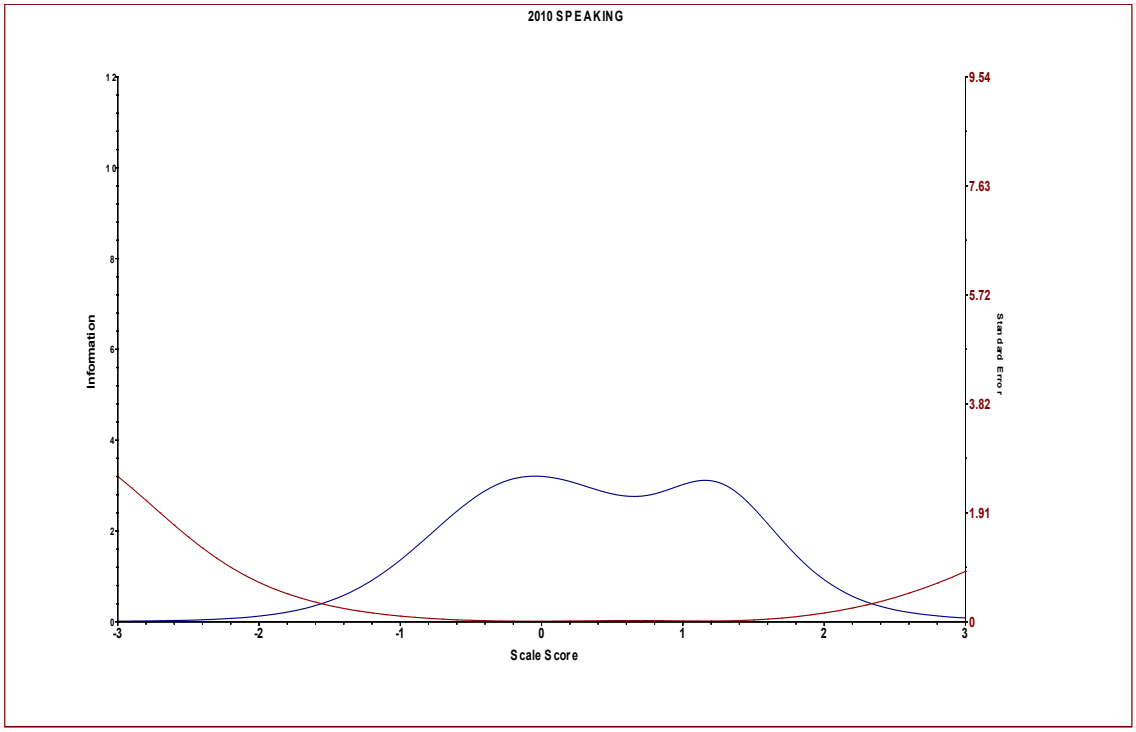
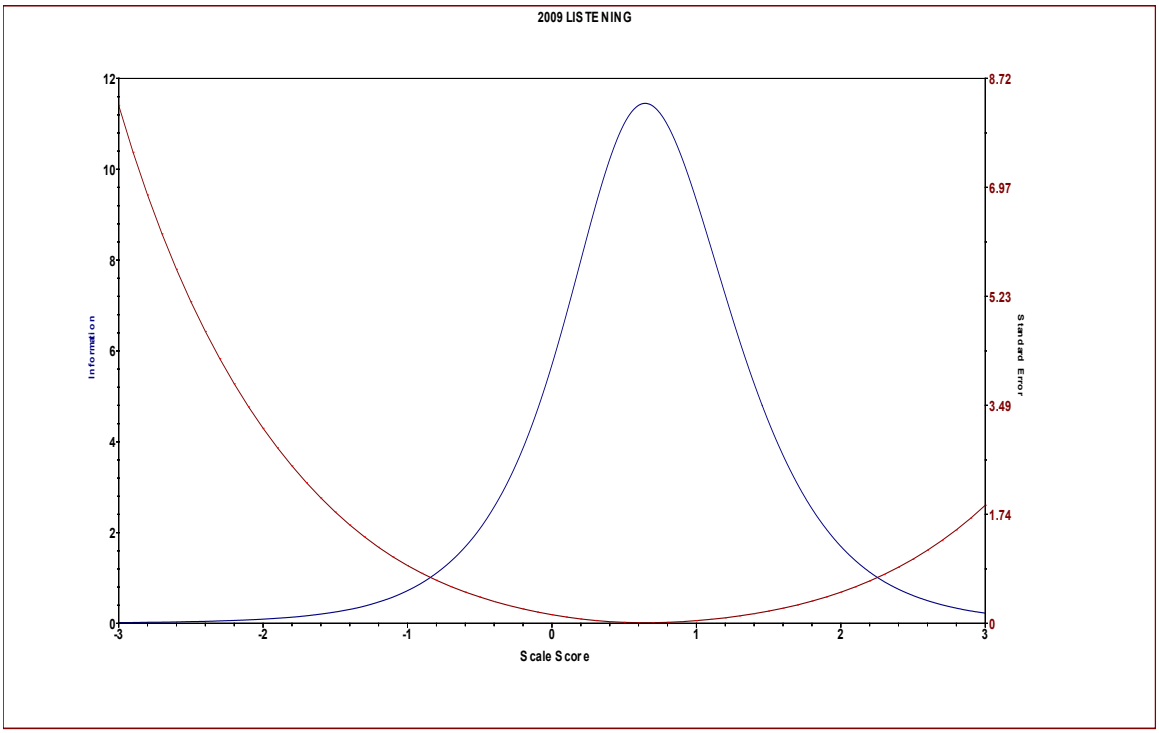


Figure 2. Test Information Functions for the 2009 and 2010 Speaking Subdomain

2009 Listening Subdomain (Number of Items = 14)



2010 Listening Subdomain (Number of Items = 15)

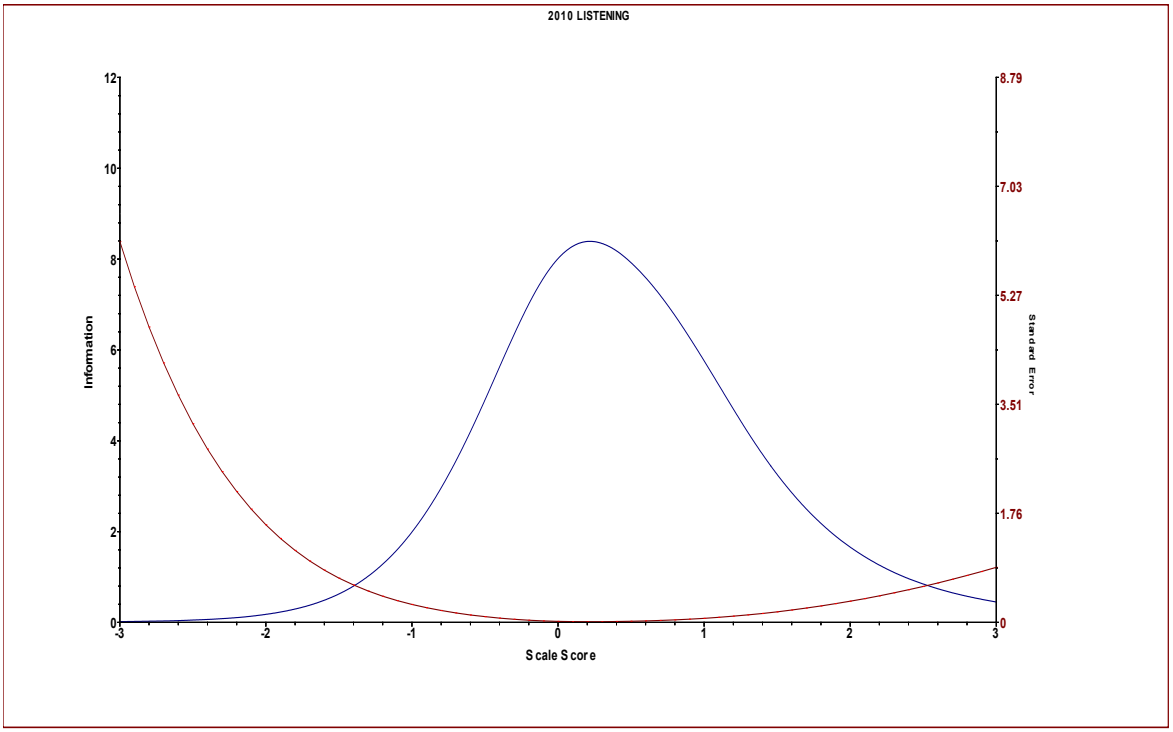


Figure 3. Test Information Functions for the 2009 and 2010 Listening Subdomain

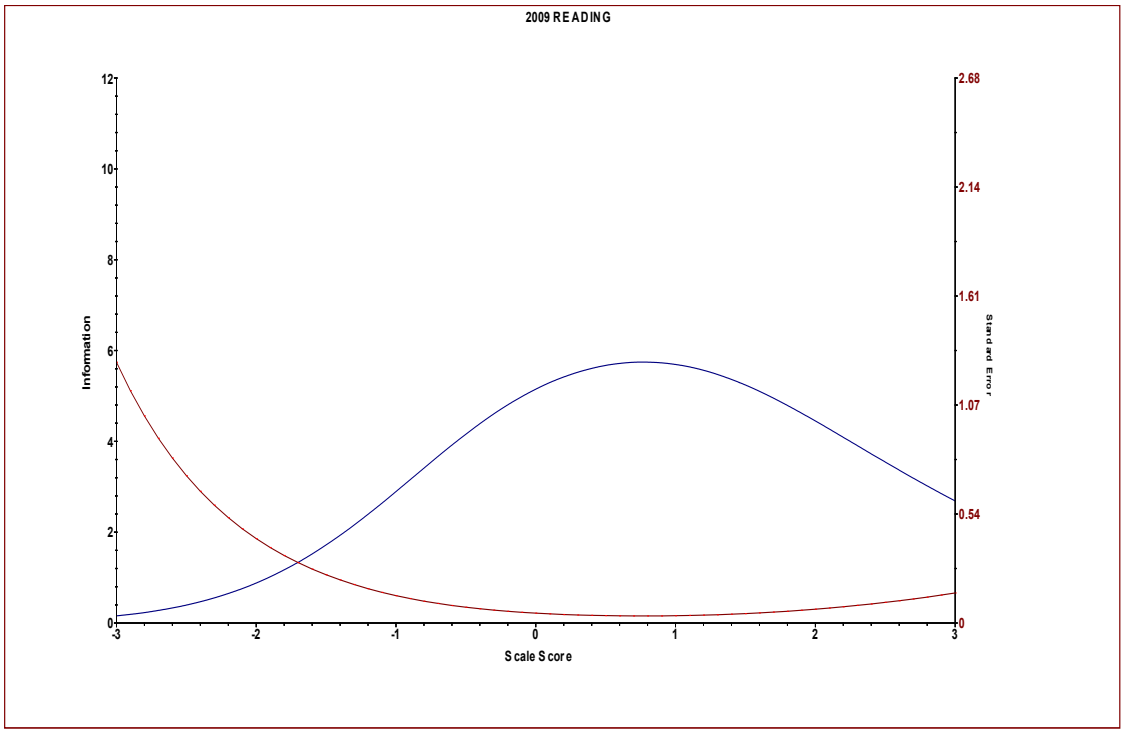
the 2010 GECAT, the number of items in the Listening subdomain in the 2010 GECAT was larger than that in the 2009 GECAT. In both years the items in the Listening subdomain, as with the Speaking subdomain, provide more precise estimates of ability for students in the above average range and are not as informative for below average students. Hence, the person ability estimates will be less precise (i.e., less reliable) for students whose ability to listen in English is below the average of the total population of examinees who responded to the test.

Test information for the Reading subdomain. Figure 4 displays the TIF for the Reading subdomain in the 2009 GECAT and another TIF for the Reading in the 2010 GECAT. Even though the Reading subdomain in the 2009 GECAT included 28 items and the 2010 GECAT included only 17 items, the test information of the 2009 GECAT was peaked at a lower level than the 2010 GECAT. This result is most likely due to the fact that the 17 items in the 2010 form are less susceptible to being answered correct by guessing than the 28 items in the 2009 forms (see Table 8). However, the 28 items in the 2009 form provide information across a broader range of student abilities.

Test information for the Writing subdomain. Figure 5 displays the TIF for the Writing subdomain in the 2009 GECAT and another TIF for the Writing in the 2010 GECAT. Even though the number of test items (80) in the 2009 GECAT was reduced to 50 in the 2010 GECAT, the number of items (5) in the Listening subdomain in the 2010 GECAT was greater than the number (4) in the 2009 GECAT. This could be why there is more information provided by the 2010 GECAT compared to the 2009 GECAT.

The test information function for the Writing subdomain in the 2010 GECAT is more peaked than the corresponding function for the 2009 GECAT, because the average IRT discrimination parameter of the 2010 GECAT (1.72) is considerably larger than the average

2009 Reading Subdomain (Number of Items = 28)



2010 Reading Subdomain (Number of Items = 18)

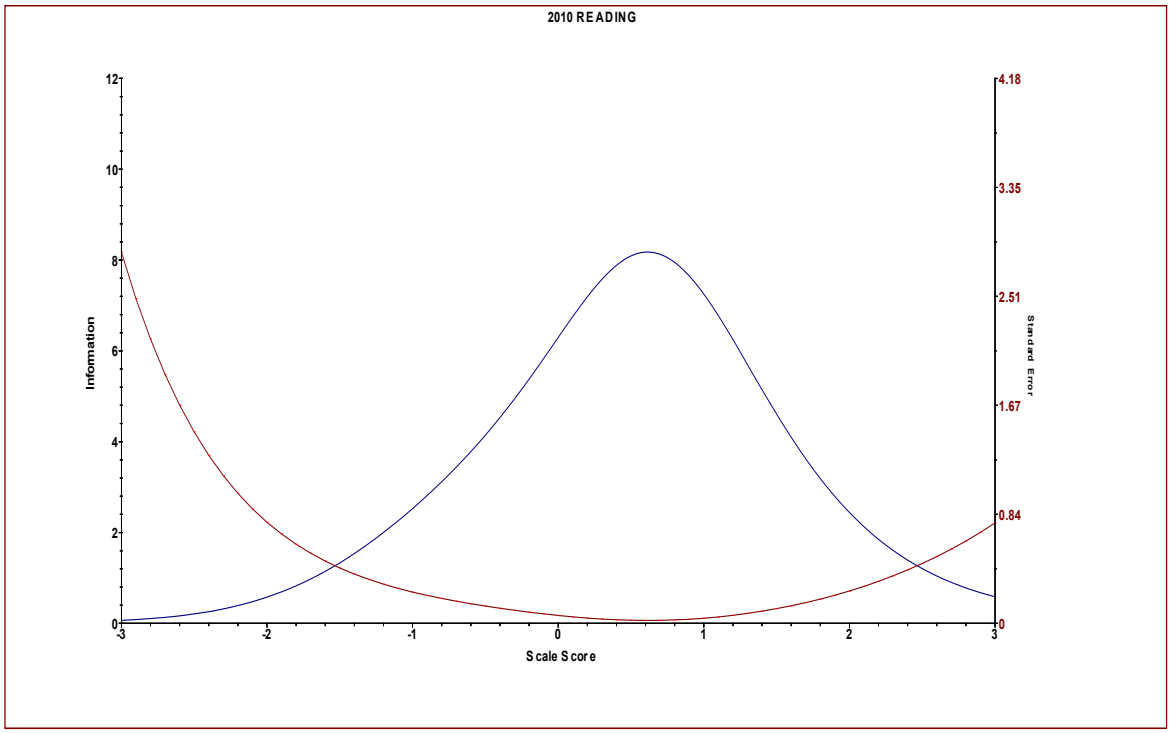
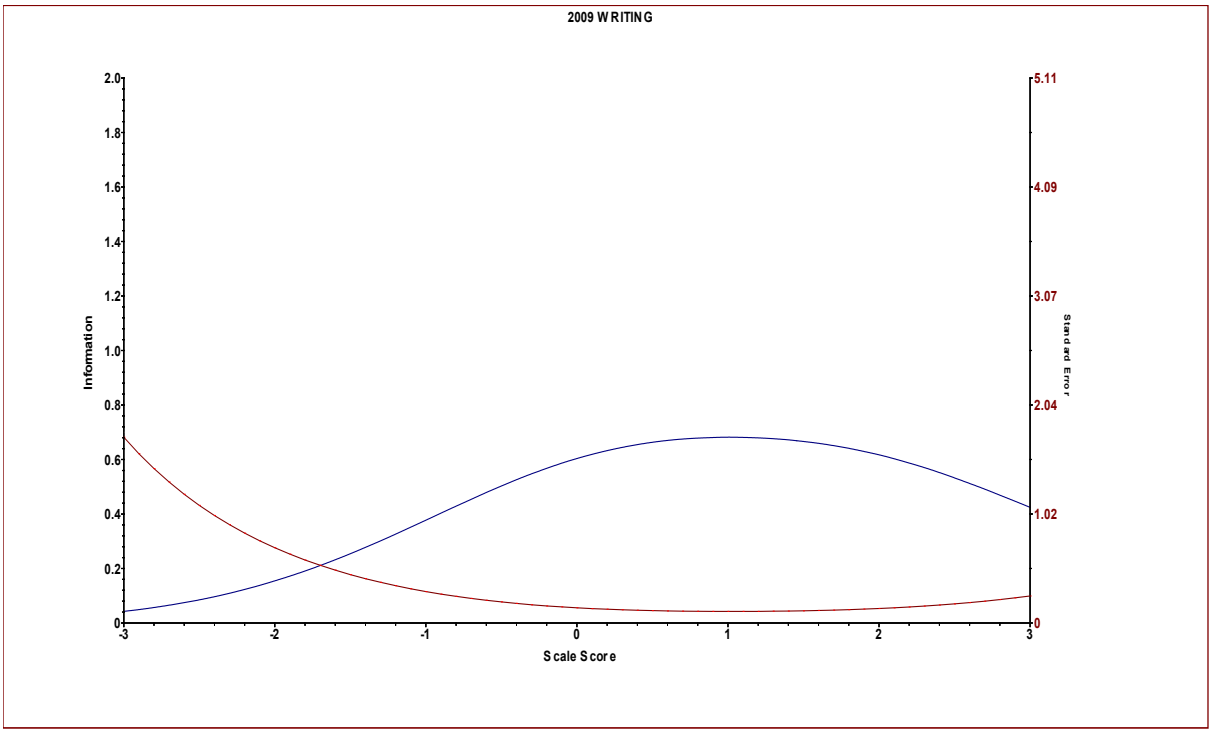


Figure 4. Test Information Functions for the 2009 and 2010 Reading Subdomain

2009 Writing Subdomain (Number of Items = 4)



2010 Writing Subdomain (Number of Items = 5)

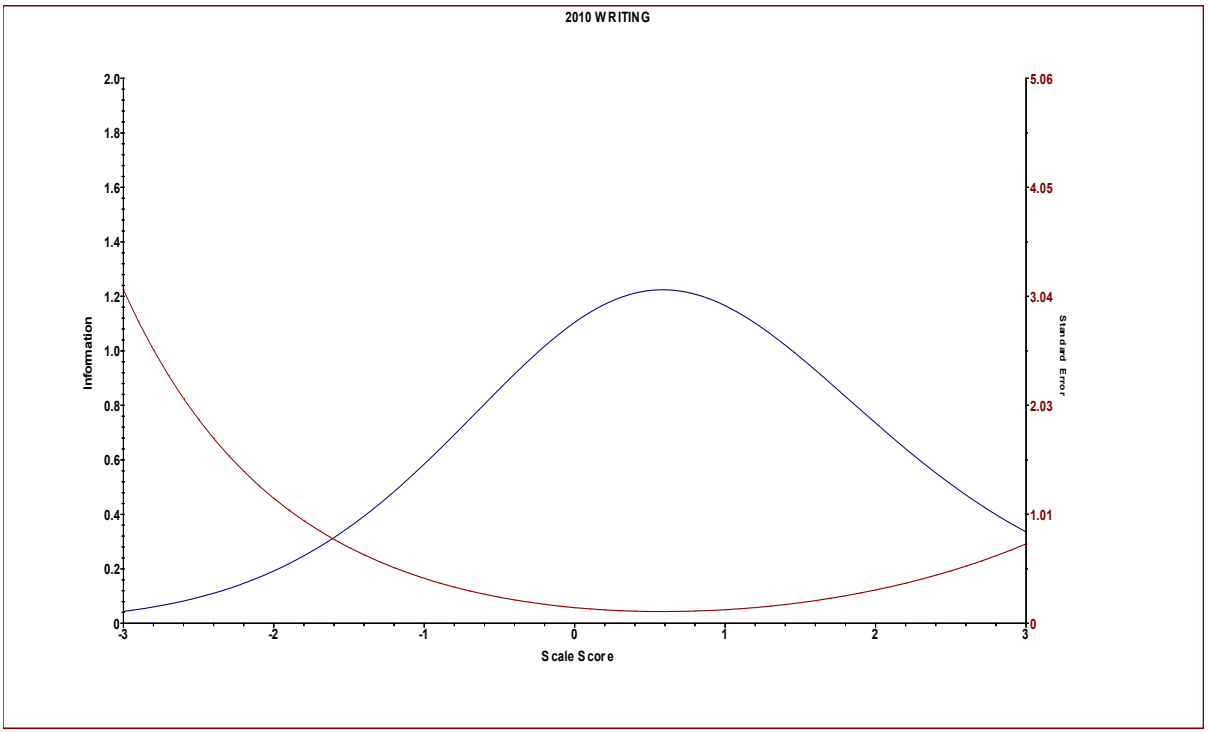


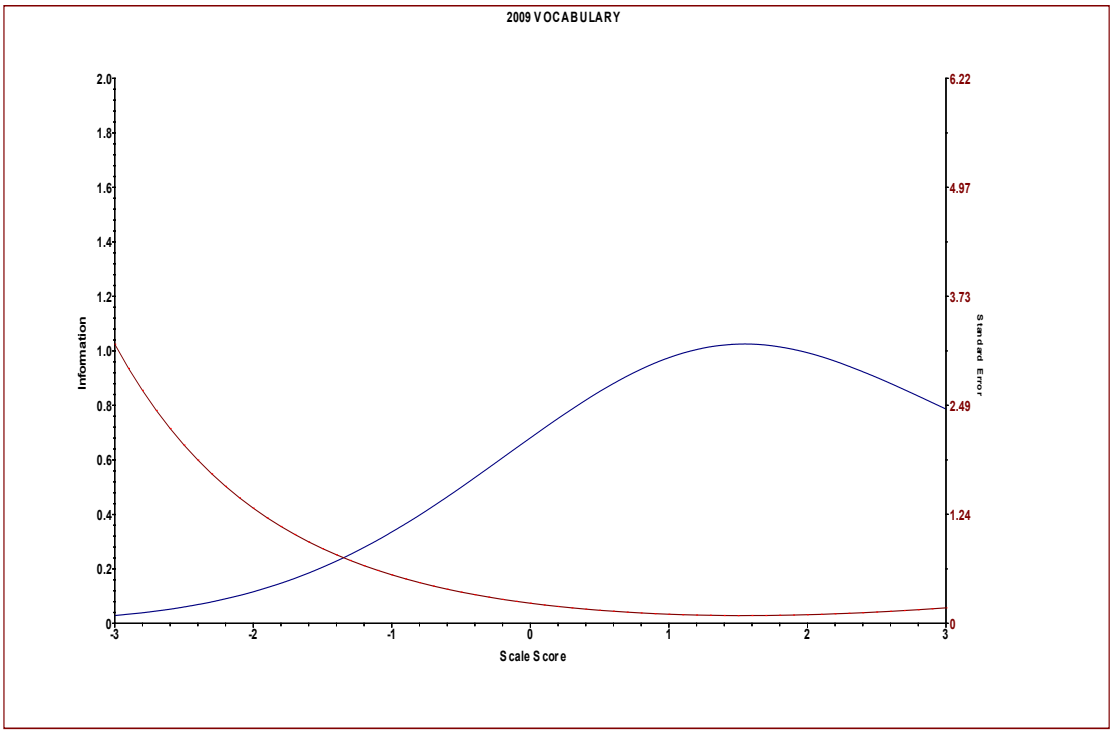
Figure 5. Test Information Functions for the 2009 and 2010 Writing Subdomain

IRT discrimination parameter (1.26) for the 2009 GECAT. In both years the items in the Writing subdomain provide more precise estimates of ability for students in the above average range, but the 2009 test information is more spread out than the TIF for the 2010 form. The 2010 form would be improved by adding a few writing items having low guessing parameters and difficulty parameters located near –1.0 or less.

Test information for the Vocabulary subdomain. Figure 6 displays a TIF for the Vocabulary subdomain in the 2009 GECAT and another TIF for the Vocabulary subdomain in the 2010 GECAT. Compared to the Vocabulary subdomain of the 2010 GECAT, the amount of information provided by the 2009 GECAT is less across all levels of ability. The skewed shape of the 2009 TIF and the location of its mode indicates that the 2009 items provide rather precise ability estimate for students who are above average. However, the Vocabulary subdomain in the 2009 version provided considerably less precise estimates for below average students.

Test information for the Grammar subdomain. Figure 7 displays a TIF for the Grammar subdomain in the 2009 GECAT. The TIF for the Grammar subdomain in the 2010 GECAT was not provided by BILOG-MG, because IRT parameter estimates of two items of the Grammar subdomain of the 2010 GECAT was not estimable in the BILOG-MG software. The function was similar in slope and the shape to that of the TIF for Vocabulary in the 2009 GECAT. This is a result of the low discrimination value of the items in the Grammar subdomain in the 2009 and 2010 GECAT. This graph shows that the items in the Grammar subdomain of the 2009 GECAT did not provide precise estimates of ability for below average students.

2009 Vocabulary Subdomain (Number of Items = 9)



2010 Vocabulary Subdomain (Number of Items = 4)

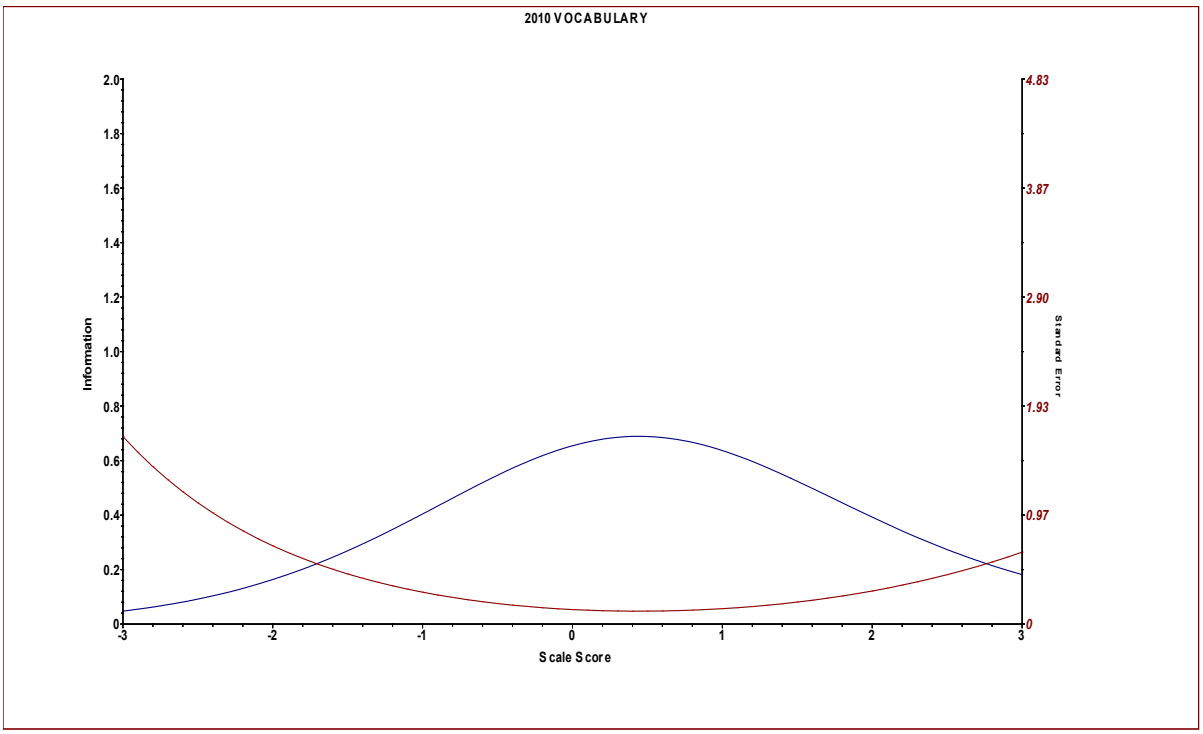
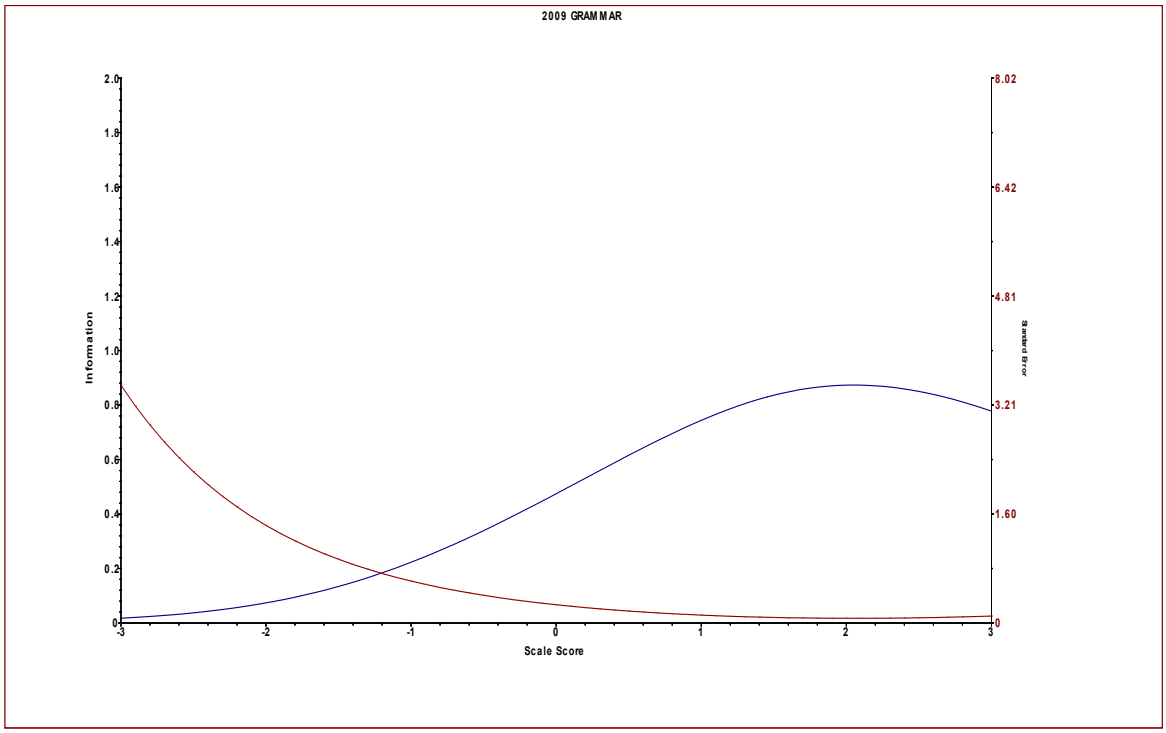


Figure 6. Test Information Functions for the 2009 and 2010 Vocabulary Subdomain

2009 Grammar Subdomain (Number of Items = 7)



2010Grammar Subdomain (Number of Items = 2)

NO TOTAL ITEM CURVES TO PLOT

Figure 7. Test Information Functions for the 2009 and 2010 Grammar Subdomain

Chapter 5: Discussion

In this study, the test items which comprise the 2009 and 2010 GECAT statistics were analyzed based on classical test theory and item response theory statistics. This chapter states the conclusions and limitations of the study, presents implications for further research, and gives recommendations for improving the GECAT.

Conclusion

Overall the quality of the test items in the 2010 GECAT seems to be better than that of the 2009 form. The combined reliability of the 2010 GECAT (.92) was identical to that of the 2009 GECAT (.92) even though the number of the test items of the 2010 GECAT decreased to 50 from the 80 test items used in the 2009 GECAT.

In general, the reliability of a test is a function of the number of the test items. That is, longer tests are generally more reliable than shorter tests. However, this generalization is true only if the items that make up the tests are similar in quality. In this study, the reliability of the scores from the 50-item 2010 GECAT are estimated to be just as reliable as scores obtained from the 80-item 2009 test.

The results of the analyses of the test items showed that both the 2009 and 2010 GECAT had weakness. First, the proportion of times allocated to each of the language subdomains varied greatly across the 2009 and 2010 forms of the GECAT. Even though the number of the test items of the 2010 GECAT decreased to 50, the number of test items in the Listening subdomain (15) was more than that of the 2009 GECAT (14), which accounted for 30% out of the total number of the test items of the 2010 GECAT. This was also true for the Writing subdomain of the 2010 GECAT. The Korea National English curriculum emphasizes the need for a balanced assessment of the four skills: listening, speaking, reading, and writing.

Another major finding was the Cronbach's Alpha reliability coefficient for the Grammar subdomain of the 2010 GECAT was only .07 while it was .23 for the 2009 GECAT. Neither of these results should be considered acceptable. In both cases, the small number of items in conjunction with the heterogeneity of the items appear to be the reasons for the low coefficients. If GOE administrators are serious about assessing students' knowledge of English grammar, then they need to increase the number of grammar items in the test.

Surprisingly, even though the 2009 and 2010 GECAT was a dichotomously scored multiple- choice test and had four options, the distracter analyses revealed that more than a few test-takers had marked Option 5 on several items. The five-option answer sheets were provided to the students at the test administration. Some students may have mistakenly marked Option 5 when they had intended to choose Option 4. But others may have marked Option 5 because they were simply marking answers without reading the item or paying attention to what they were doing. Ideally, answer sheets would be prepared to match the number of options presented with each item used in the GECAT.

Another finding dealt with one of the main objectives of the administration of the GECAT that is to increase students' motivation to learn English. Total information functions of six subdomain tests (speaking, listening, reading, writing, vocabulary and grammar) showed that most of the test information functions of the 2009 GECAT were peaked at ability levels in the range of $0.9 < \theta < 1.5$, while those of the 2010 GECAT were peaked at ability levels in the range of $0.0 < \theta < 0.6$. This change seems to be an improvement, but such test items of the GECAT may frustrate the drive for the students to learn English to the students below the intermediate ability level of $\theta = 0.0$. In consequence, it might be beneficial to increase the number of the easy test items for the student below the intermediate ability level

of theta =0.0 which represents average performance.

Limitations

Two limitations exist in this study and must be acknowledged. One limitation concerns unidimensionality. The other limitation refers to the lack of generalizability of the findings.

Unidimensionality assumption. The dimensionality of the 2009 and 2010 forms of the GECAT was not formally investigated in this study. One way to empirically assess the dimensionality of scores obtained from each of the GECAT forms would have been to use factor analysis. However, that was beyond the scope of this masters thesis and was not done.

The researcher in this study concluded that the assumption of unidimensionality is not likely to be defensible in the context of a 50- or 80-item test designed to simultaneously assess multiple language modalities. Students' knowledge and skills in the six language modalities probably are not discrete competencies that exist independently of each other. On the other hand, they are also not likely to constitute a simple, unidimensional trait. They are more likely to be a complex set of intercorrelated skills and capacities with some of them being more highly interrelated than others.

Furthermore, unlike tests in other subject-matter areas, language tests make use of the very language that is the target of the assessment to function as a tool for administering the test and communicating the tasks to be tested to the examinees. Therefore, to the degree to which individual examinees lack reading ability their performance on other sections of the GECAT designed to assess other language skills may be impaired.

Consequently, the assumption of unidimensionality is arguably not defensible in the context of the present study. For this reason, the items in both the 2009 and 2010 forms of the GECAT were subdivided into six different subdomains that were each analyzed as

separate unidimensional traits. The IRT parameter estimates resulting from these six separate IRT analyses were then transformed to a common metric in order to make them commensurable.

Lack of generalizability. A second limitation of the study is the results of the analysis of the test items may not be generalized to all the situations because the GECAT test is a test tool developed for students learning English as a foreign language. These students tend to learn English only in the classroom and don't have many opportunities to converse or make themselves understood in English.

Implications for Further Research

The results of this study point to three implications for further research. First, a series of studies focused on validity issues should be conducted. The present study does not address validity concerns. However, the decision to reduce the GECAT from 80 items to 50 items raises a number of important issues directly related to content validity and indirectly related to construct validity. Both the 2009 and 2010 forms of the GECAT included multiple-choice items sampled from the same six subdomains representing different aspects of English language knowledge and usage. However, the relative emphasis given to sampling test items from each of these subdomains changed considerably from 2009 to 2010. This change in emphasis is displayed in Table 3. The relative emphasis given to the Reading subdomain changed very little. It decreased from 35 percent in 2009 to 34 percent in the 2010 form. However, the relative emphasis given to assessing Speaking decreased from 23 percent in 2009 to 14 percent in 2010, while the percent relevant to assessing Listening increased from 18 percent to 30 percent. At the same time, the emphasis given to assessing Writing increased from 5 to 10 percent, while the emphasis given to assessing Vocabulary knowledge decreased from 11 to 8 percent and the emphasis given to assessing grammar usage decreased

from 9 to 4 percent.

A second implication for research is for the GOE. They should consider developing and administering more direct assessments of students' proficiency in Writing and Speaking. Writing and speaking are productive skills, but selected response test items (e.g. as multiple-choice items) are more appropriate for assessing receptive skills or knowledge such as Vocabulary and Grammar usage. If GOE administrators and other educational leaders believe that it is important to assess students' proficiency in writing and speaking in English, then they should consider replacing the Writing and Speaking sections of the GECAT with performance assessments such as an essay test for assessing writing and some way of collecting spoken responses in order to assess students' Speaking competence. Such direct assessments would be much more time consuming and expensive to administer and rate, but if properly developed and administered, they would likely produce more valid estimates of students strengths and weaknesses in using these productive language skills.

One way to use the more direct approach to assessing Writing and Speaking would be implement a sampling plan where a direct Writing assessment would be administered to a random half of the students, and a direct speaking assessment would be administered to the other half.

In the process of developing these direct assessments, it would be helpful to test a sample of students with both the multiple-choice writing items and a direct writing assessment. Such a study would permit GOE to determine how well scores on the multiple-choice version of the writing test predict students' performance on the direct Writing assessment. This same approach could be used with direct and indirect forms of the Speaking assessment.

Another direction for research should focus on the validity of the GECAT scores. A series of factor analysis studies—including both exploratory and confirmatory factor analysis—should be conducted to obtain evidence of construct validity of scores from the GECAT. Both the 2009 and 2010 data sets should be included in these analyses to ascertain how the factor structure of the GECAT changed when the number of items was decreased from 80 to 50.

Recommendations for Improving Future GECAT Forms

This study provided much information that can be used to make the GECAT better. The following five recommendations for improving the GECAT and its usage are based on the analysis and interpretation of the findings of this study.

The initial recommendation is that the GOE administrators should consider abandoning the process of generating a completely new set of multiple-choice items for the GECAT every year and replacing this start-over-every-year procedure with an item banking process that involves carefully defining a table of specifications, writing items to sample each aspect of that table, and then having each proposed item reviewed and screened by a knowledgeable committee of reviewers and then pilot-tested on a reasonable sample of representative examinees. Writing draft versions of multiple-choice items and then having them reviewed, screened, pilot-tested, and subsequently revised is too demanding, time consuming, and expensive to repeat every year, and does not permit GOE to systematically accumulate over the years a pool of high-quality items that have been sufficiently scrutinized and refined.

Next, students' responses to the GECAT should be item analyzed every year. Ideally, IRT would be used to conduct these analyses. At a minimum, CTT difficulty and discrimination statistics should be computed for every item. In addition, distracter analyses should be performed for each item that has a low adjusted-total correlation coefficient, and reliability

coefficient should be computed. The results should be used as a basis for improving the GECAT items over time.

GECAT scores are used as a basis for assigning grades to students. Because of its importance, GOE should consider developing parallel forms of the test and conducting a formal equating study so that scores from the two forms would be exchangeable.

Another recommendation is that instead of arbitrarily using 90% or above as the criterion for receiving an A, 80% to 89% for a B grade, and 70-79% for a C grade, GOE should also consider conducting a standard setting study to determine the appropriate cut-scores as a basis for making these grading decisions. Such a study is especially important since students who pass the test with a C or higher grade are awarded a certificate of English language proficiency.

Finally, for long-range purposes, GOE should consider creating a computer-adaptive version of the GECAT. A computer-adaptive form may not be feasible at present because of a lack of sufficient computer hardware in each school. As the schools accumulate more and better computers, a computer-adaptive GECAT will become more practical in the future. A computer-adaptive version of the GECAT would likely be more efficient in terms of testing time than static paper-and-pencil versions because a precise estimate of students' proficiency can be obtained with fewer items when the items are selected to fit the each examinee's general ability level.

REFERENCES

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

Baker, F.B. (1978). Advances in item analysis. *Review of Educational Research*, *47*, 151-178.

Brown, H.D. (2000). *Principles of Language Learning and Teaching(4th ed.)*. NY: Pearson Education.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*, 1-47.

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

de Ayala, R.J. (2009). *The theory and practice of item response theory*. New York: Guilford.

Demars. C. (2010). *Item Response Theory*. London: Oxford University Press.

Ebel, R.L & Frisbie, D.A. (1986). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hambleton, R.K. & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *An NCME Instructional Module(16)*. National Council on Measurement in Education. *Retrieved from* http://www.ncme.org/pubs/items/24.pdf.

Hambleton, R.K., (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp.147-200). New York: Macmillan.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130-150.) Washington, DC: American Council on Education.

Hymes, D.H. (1972). On communicative competence. In J.B. Pride and J. Holmes (Eds.), *Sociolinguistics.* (pp. 269-293). Harmondsworth: Penguin.

Livingston, S.A. (2006). Item analysis. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 421-441). Mahwah, NJ: Erlbaum.

Lord, F.M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

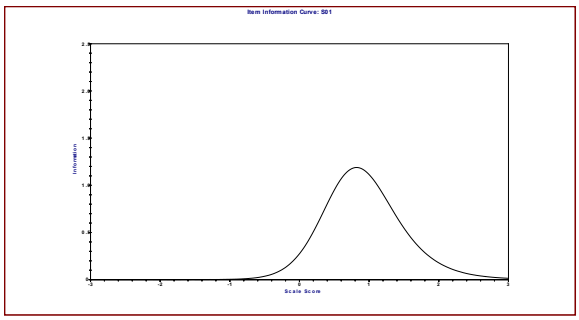
Oosterhof, A.C. (2001). *Classroom applications of educational measurement* (3rd ed.) Upper Saddle River, NJ: Merrill Prentice-Hall.

Reynolds, C.R., Livingston, R.B. & Willson, V. (2010). *Measurement and Assessment in Education* (2nd ed.). NJ: Pearson Prentice-Hall.

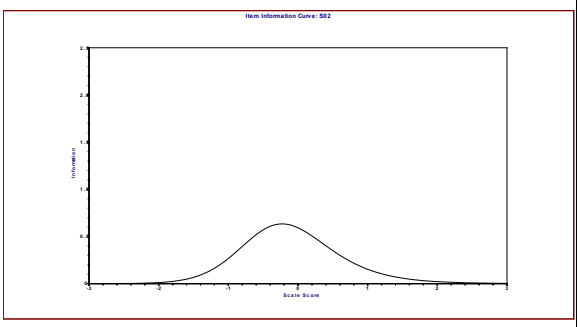
Yen, W. M. & Fitzpatrick, A.R. (2006). Item response theory. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). New York: Macmillan.

Appendix A. Item Information Functions for the Speaking subdomain in the 2009 GECAT

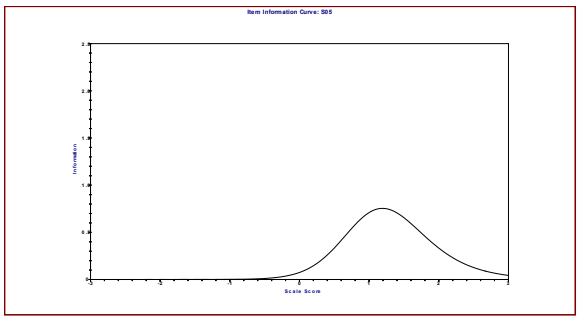
Item 01



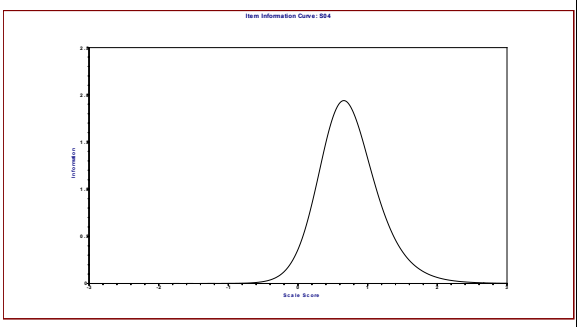
Item 02



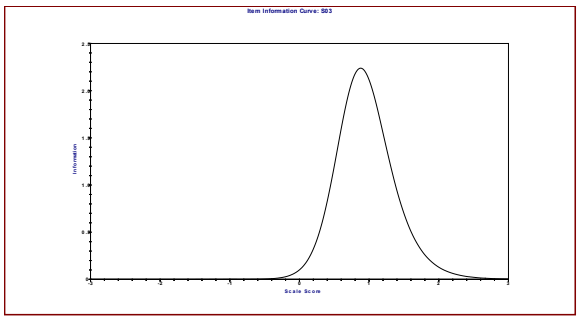
Item 03



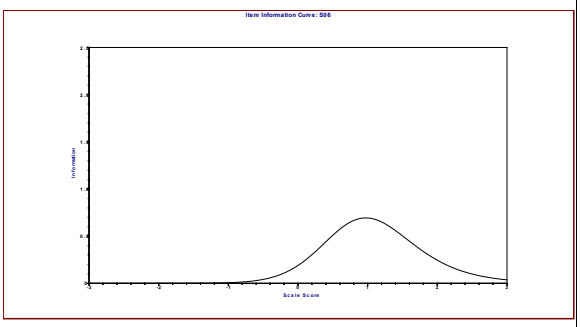
Item 04



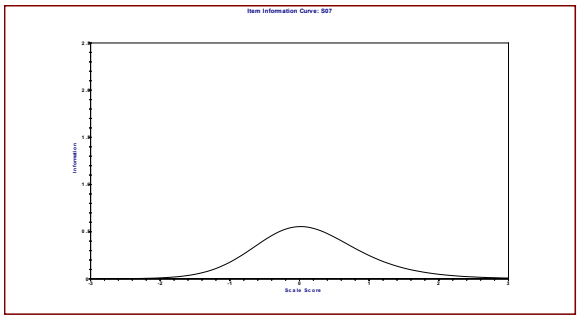
Item 05



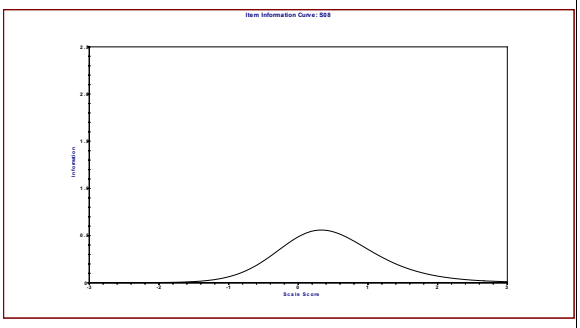
Item 06



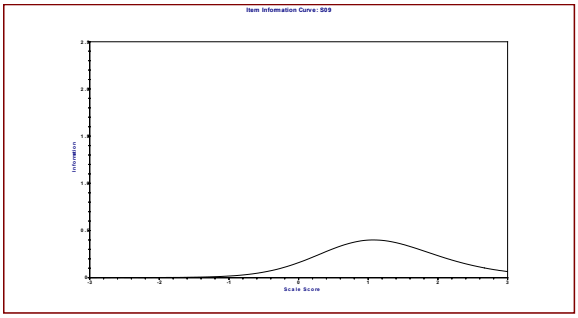
Item 07



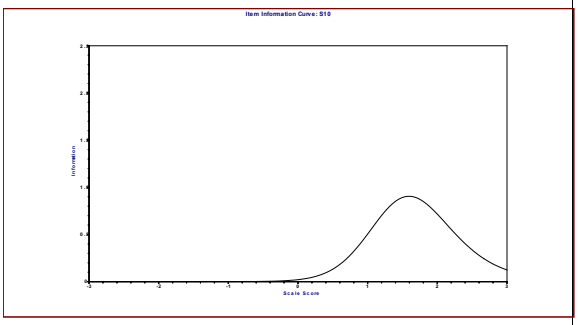
Item 08



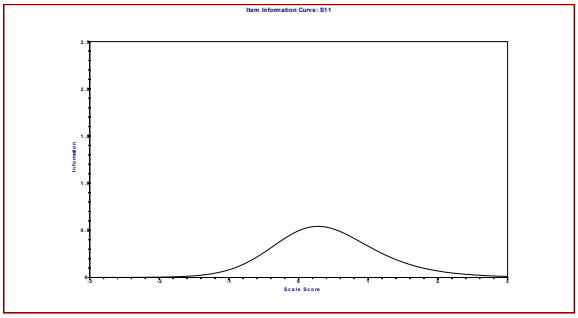
Item 09



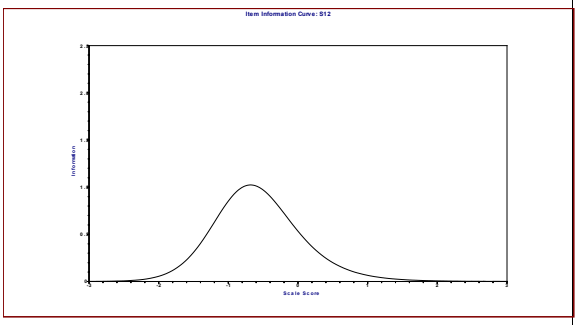
Item 10



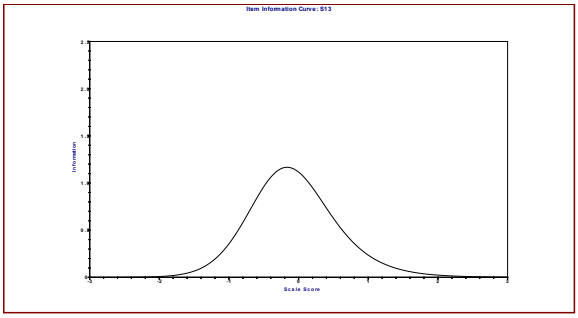
Item 11



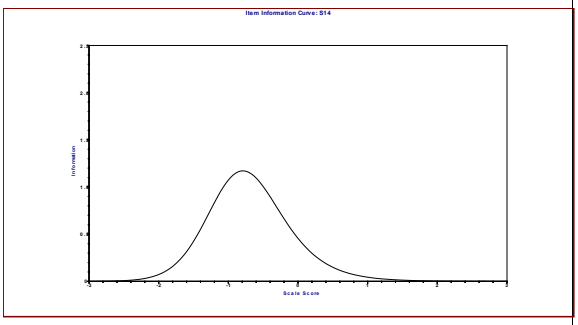
Item 12



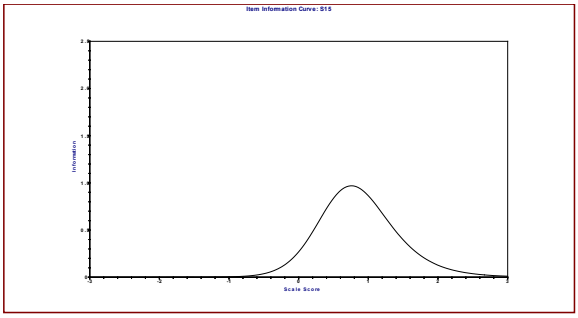
Item 13



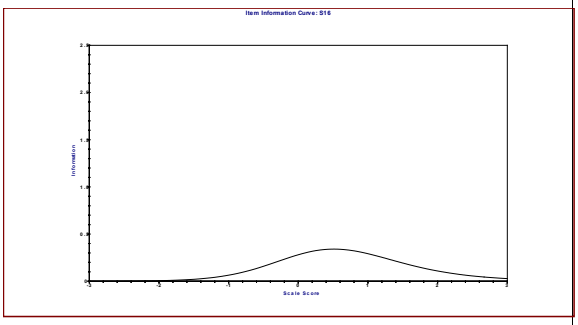
Item 14



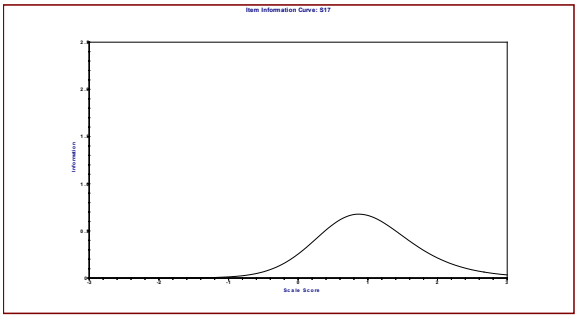
Item 15



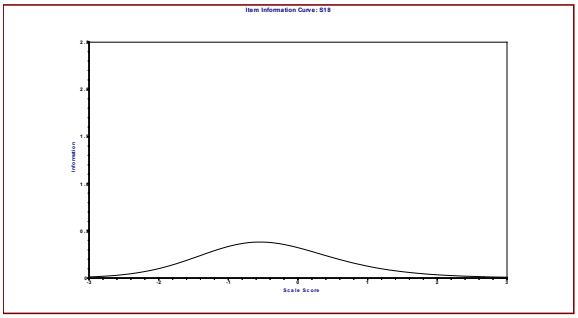
Item 16



Item 17

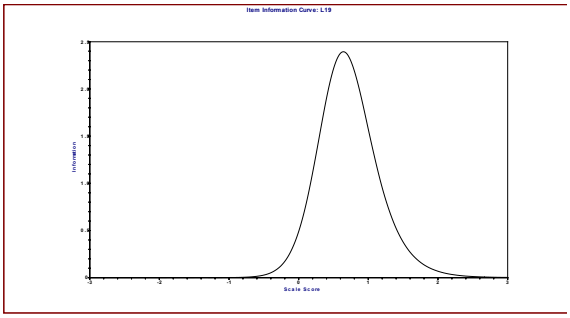


Item 18

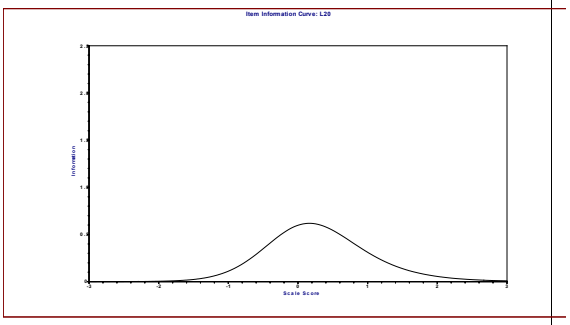


Appendix B. Item Information Functions in Listening Subdomain in the 2009 GECAT

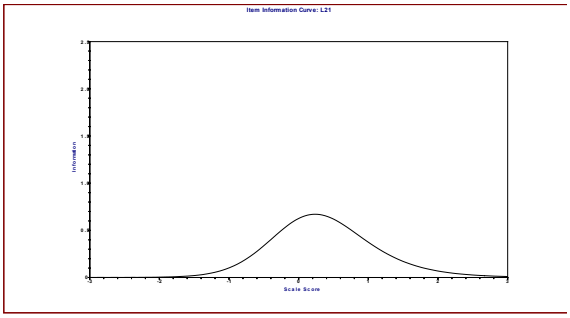
Item 19



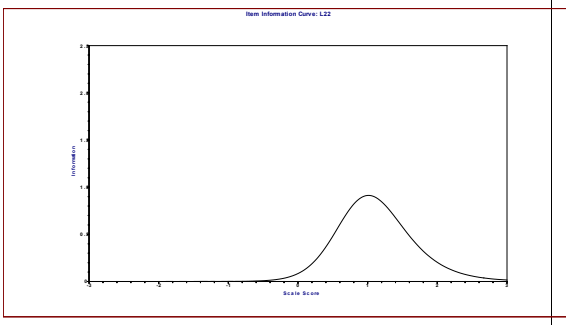
Item 20



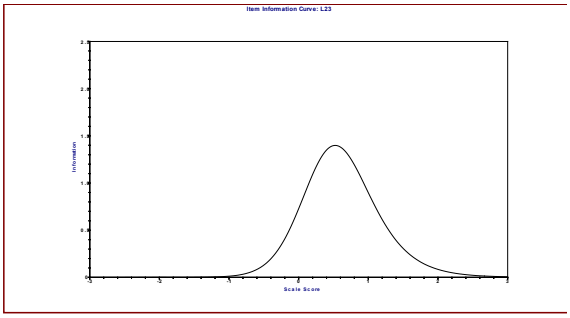
Item 21



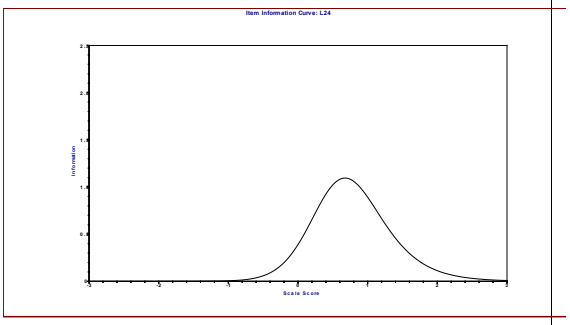
Item 22



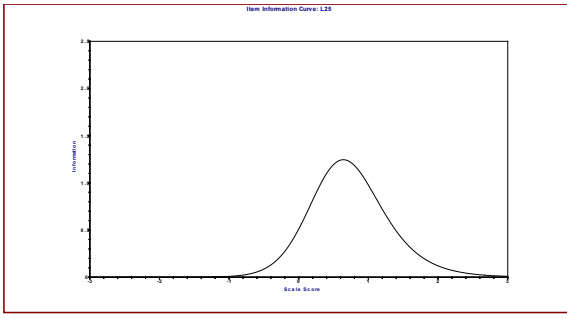
Item 23



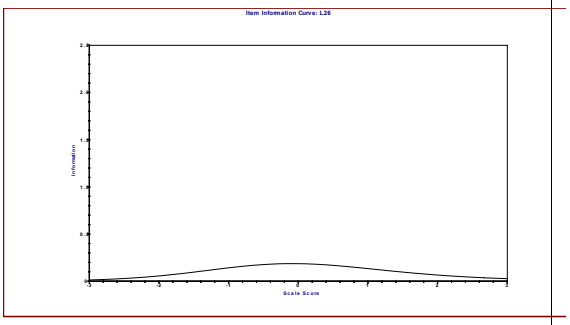
Item 24



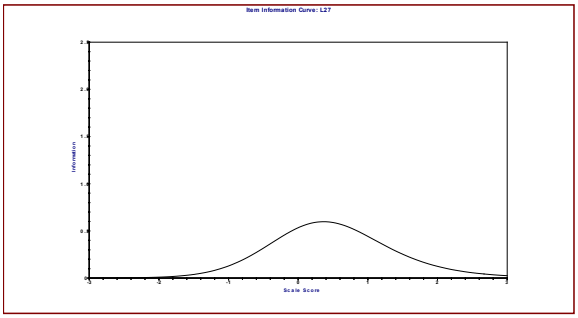
Item 25



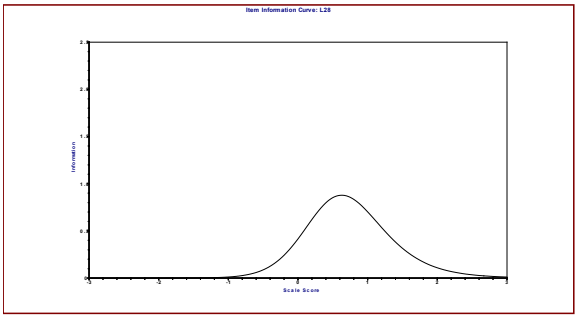
Item 26



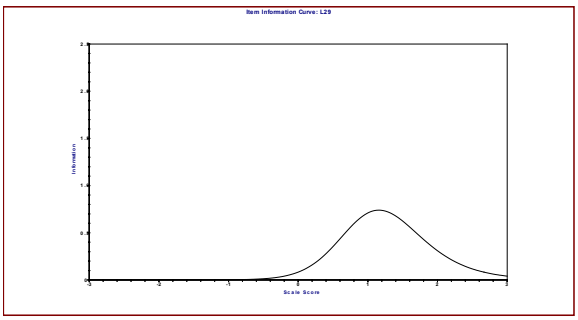
Item 27



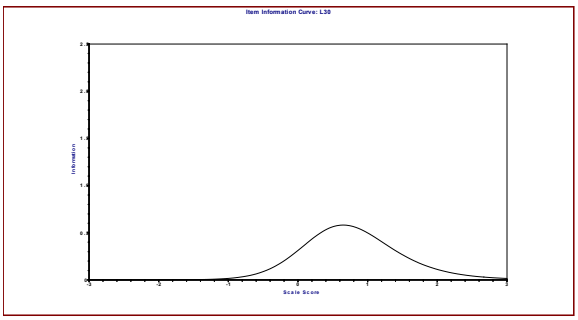
Item 28



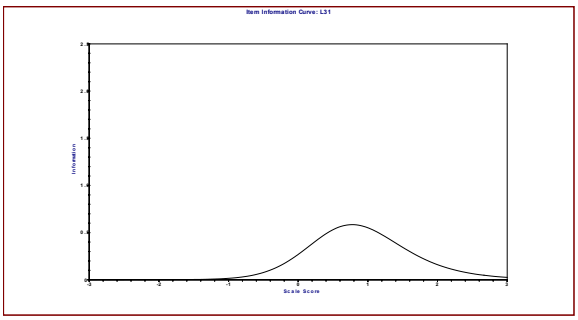
Item 29



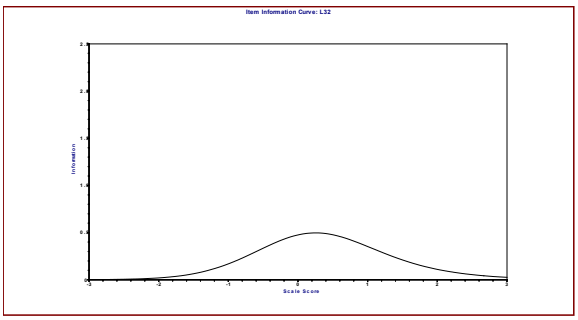
Item 30



Item 31

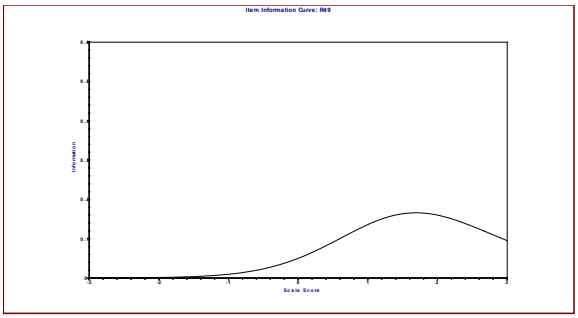


Item 32

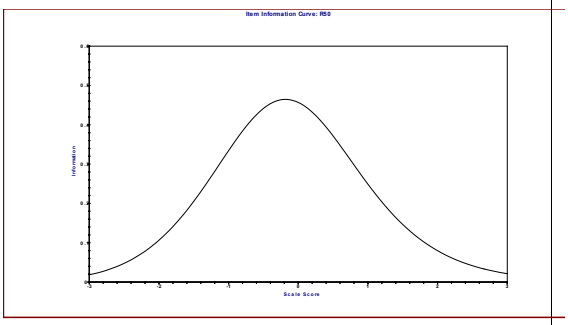


Appendix C. Item Information Functions for the Reading subdomain in the 2009 GECAT

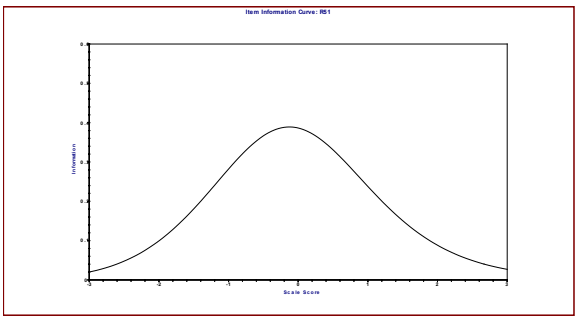
Item 49



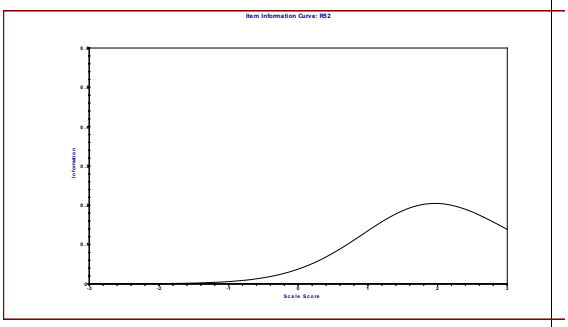
Item 50



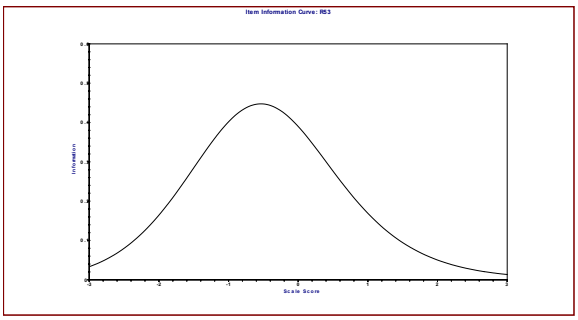
Item 51



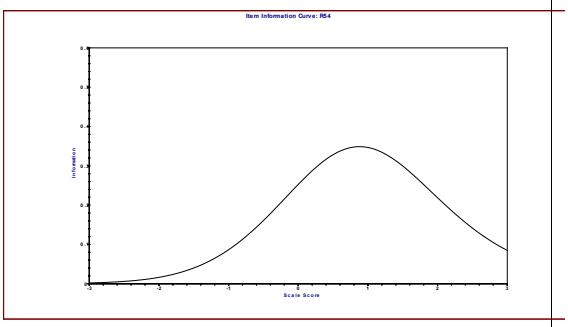
Item 52



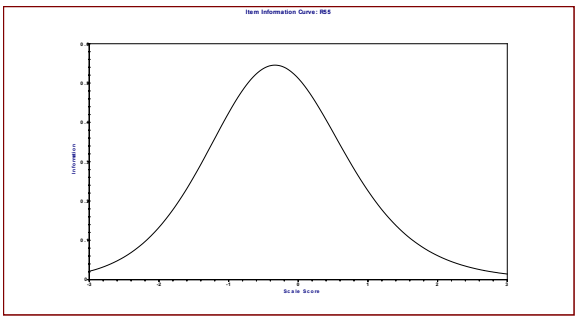
Item 53



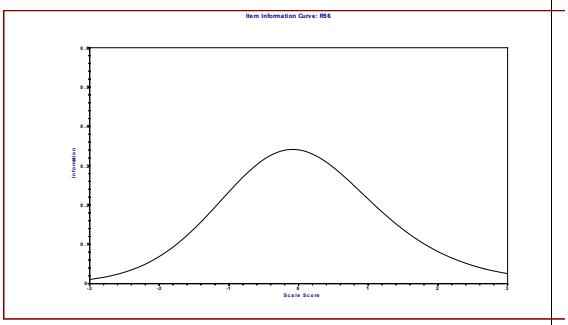
Item 54



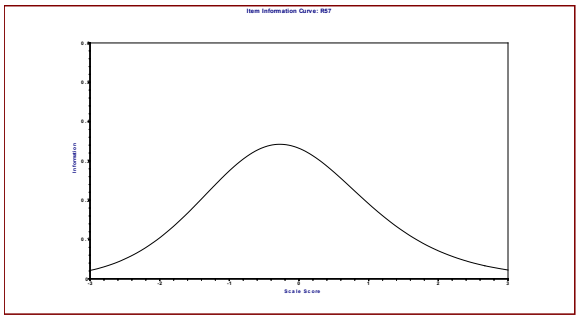
Item 55



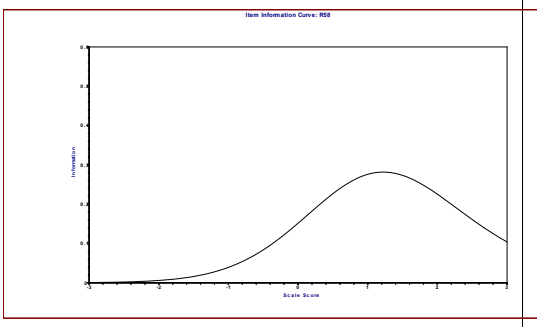
Item 56



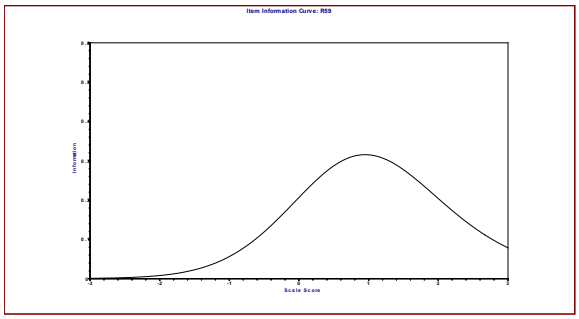
Item 57



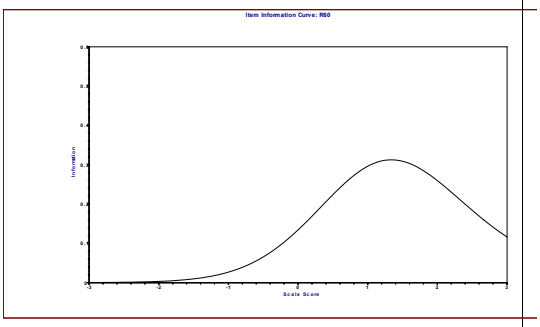
Item 58



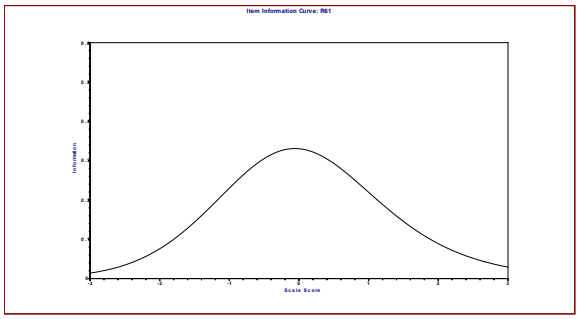
Item 59



Item 60



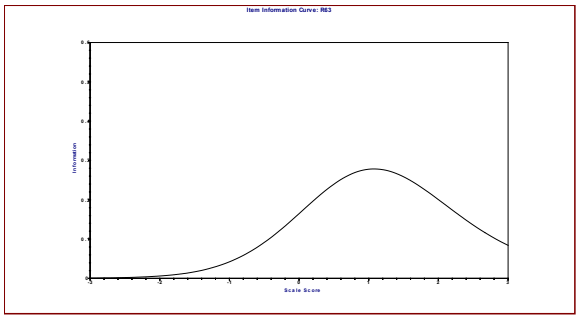
Item 61



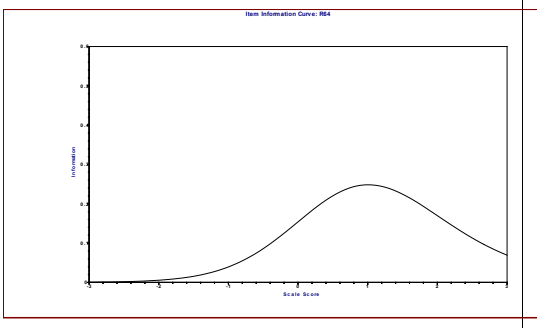
Item 62

Not Estimable

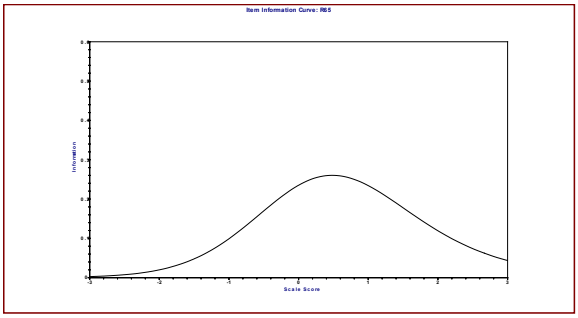
Item 63



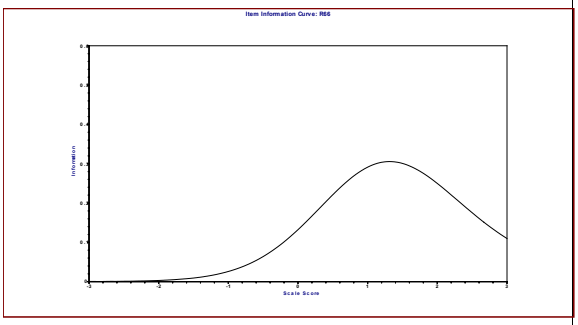
Item 67



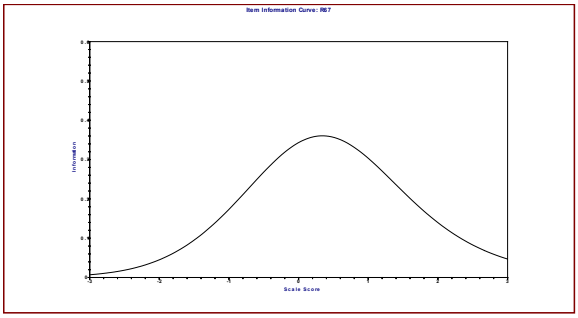
Item68



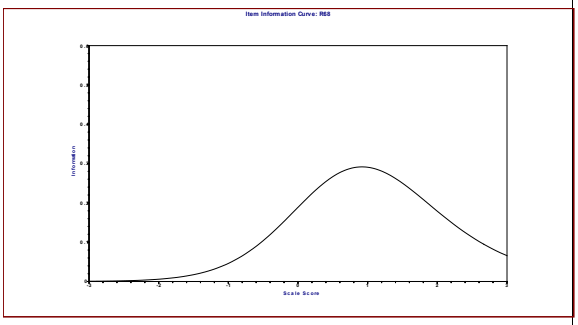
Item 69



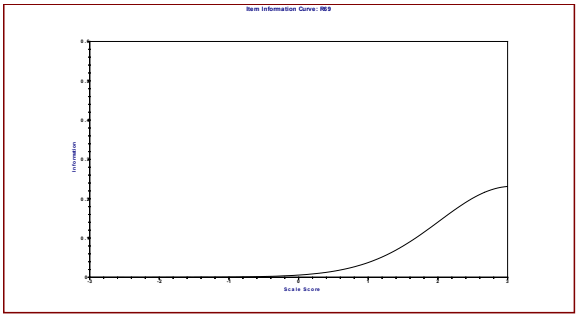
Item 70



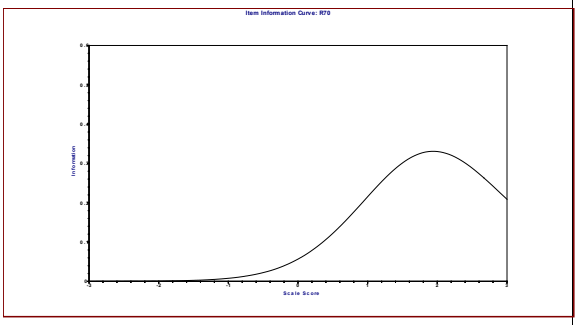
Item 71



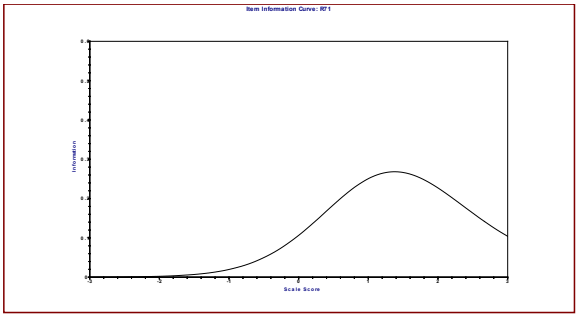
Item 72



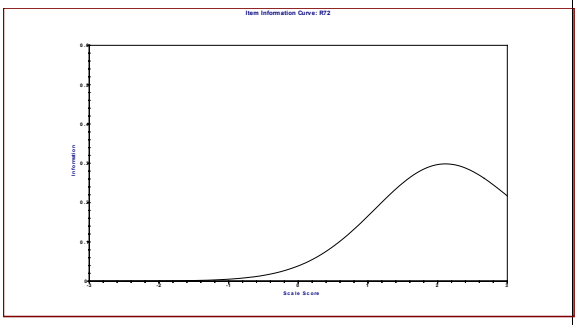
Item 73



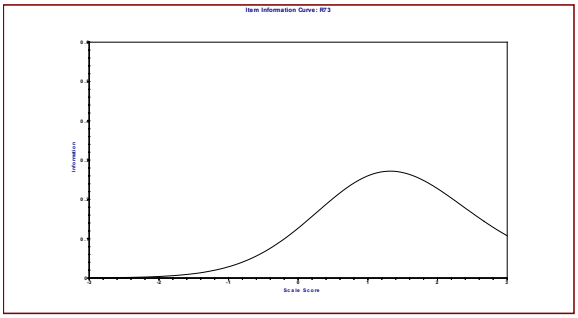
Item 74



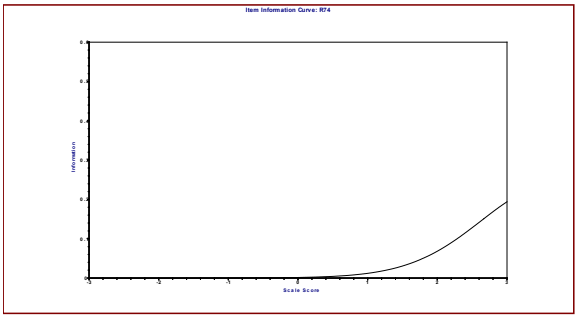
Item 75



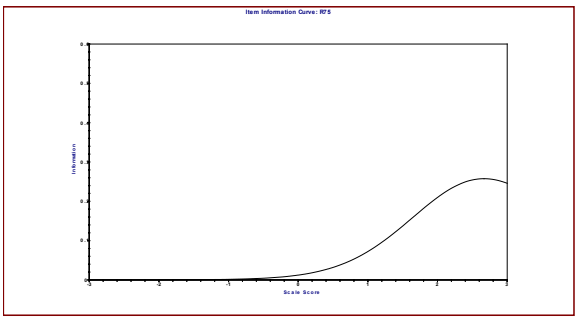
Item 76



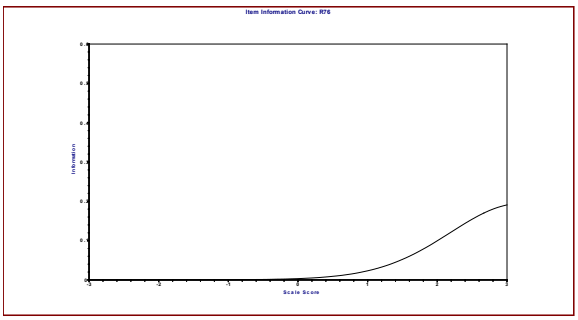
Item 77



Item 78

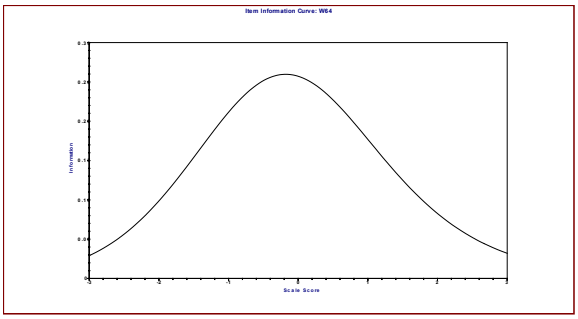


Item 79

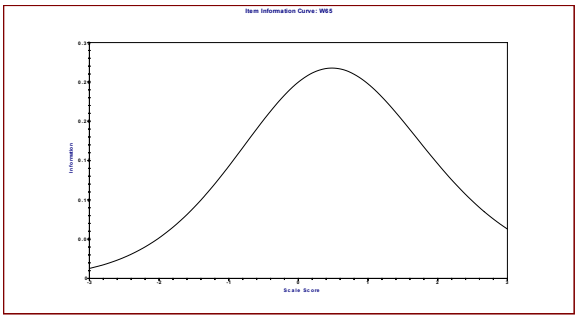


Appendix D. Item Information Functions for the Writing subdomain in the 2009 GECAT

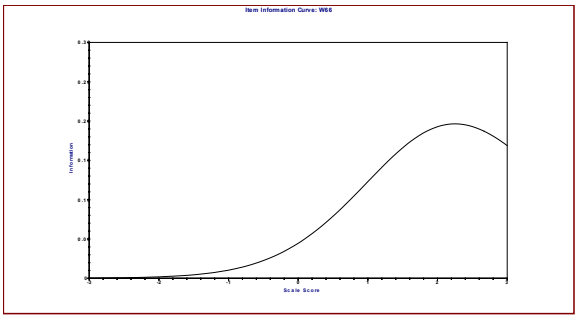
Item 64



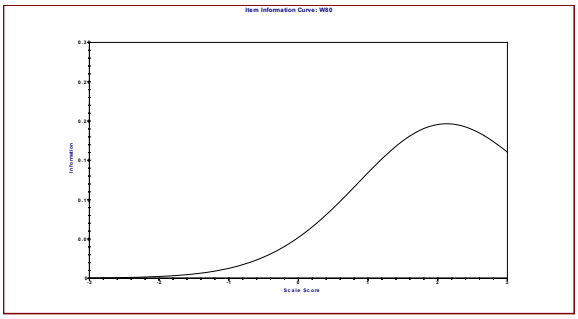
Item 65



Item 66

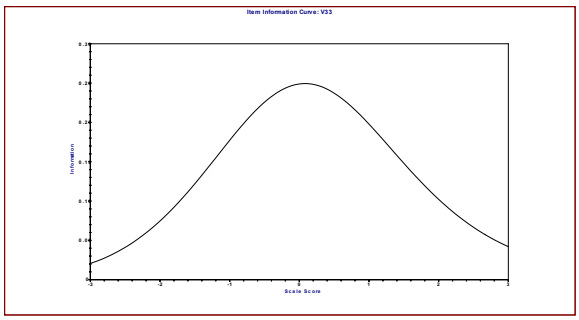


Item 80

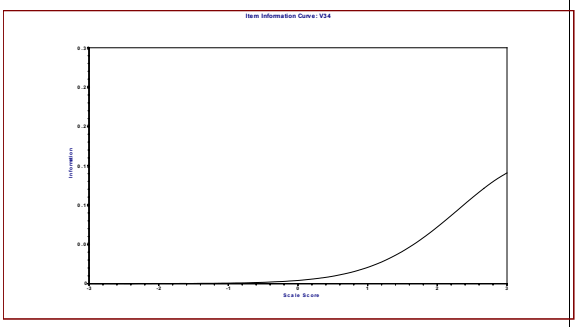


Appendix E. Item Information Functions for the Vocabulary subdomain in the 2009 GECAT

Item 33



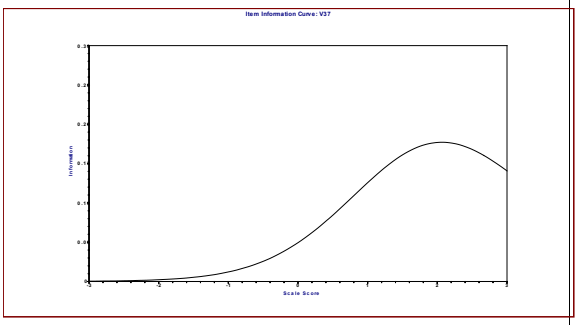
Item 34



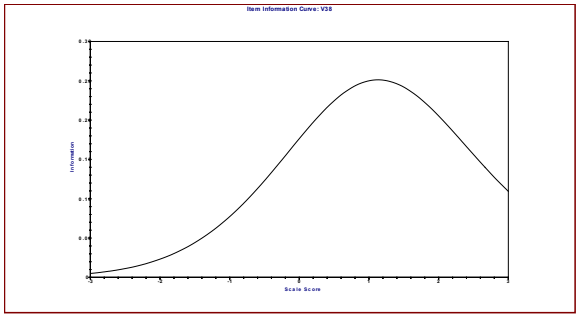
Item 35, Item 36

Not Estimable

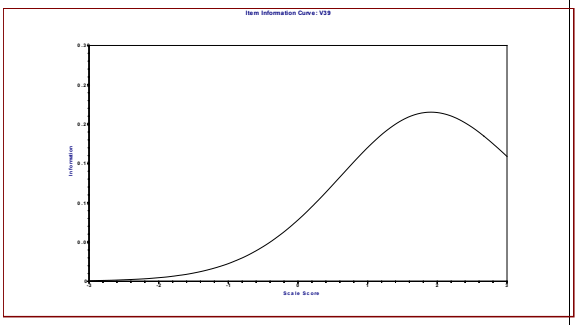
Item 37



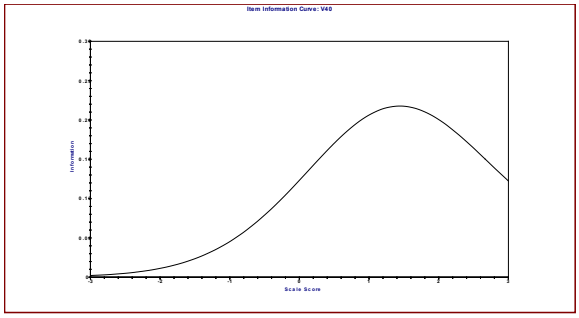
Item 38



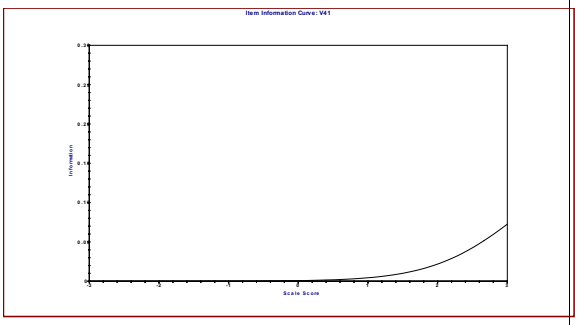
Item 39



Item 40

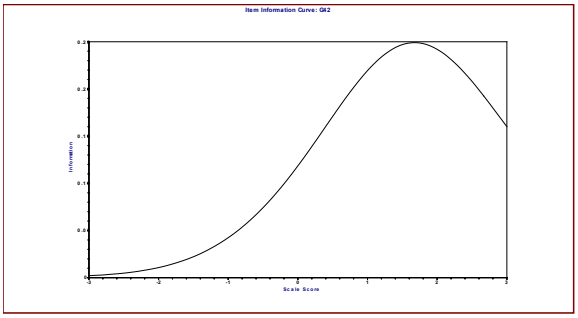


Item 41

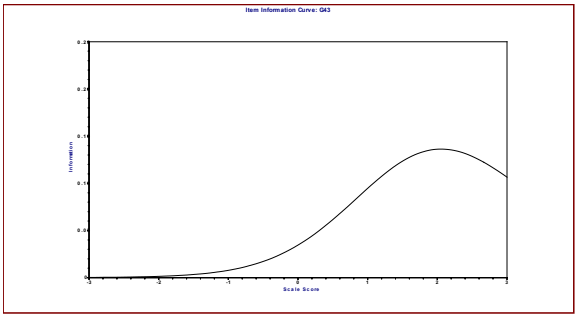


Appendix F. Item Information Functions for the Grammar Subdomain in the 2009 GECAT

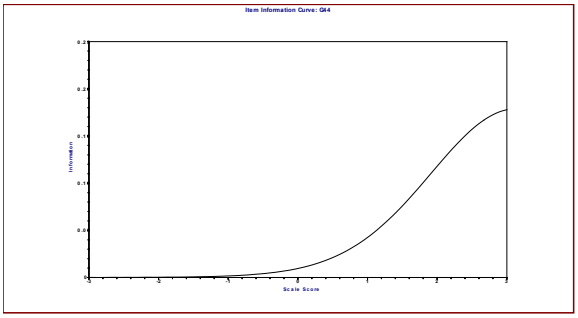
Item 42



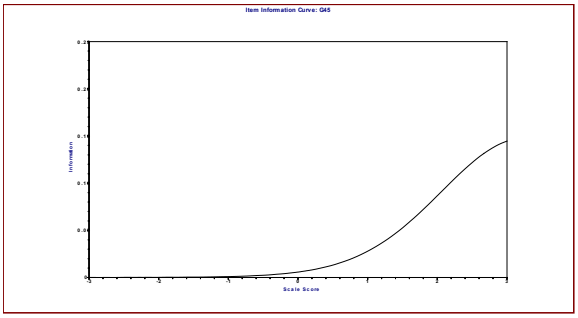
Item 43



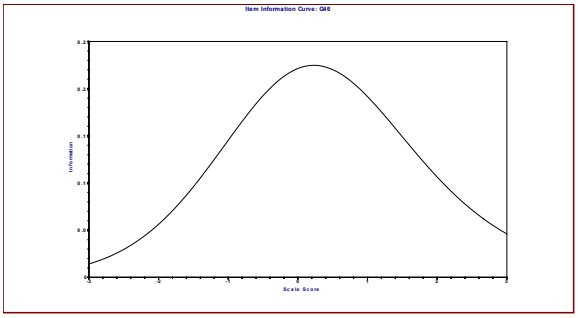
Item 44



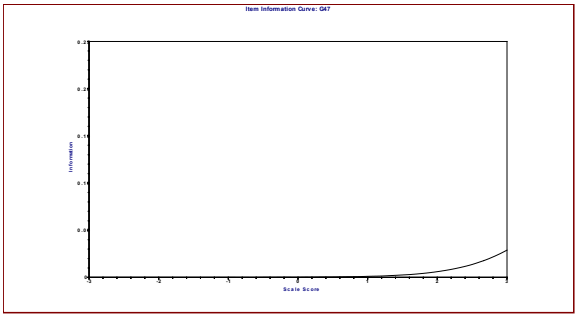
Item 45



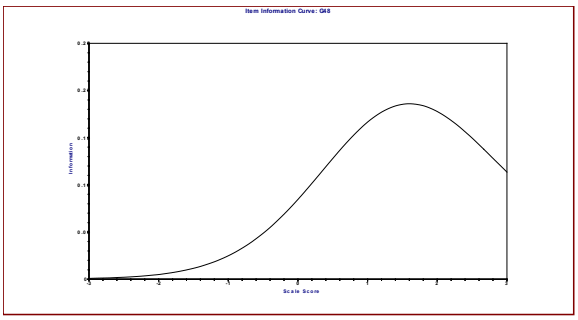
Item 46



Item 47

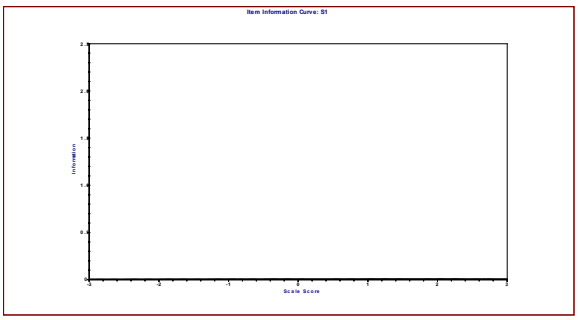


Item 48

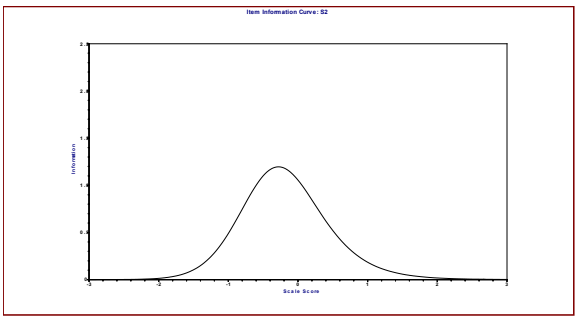


Appendix G. Item Information Functions for the Speaking subdomain in the 2010 GECAT

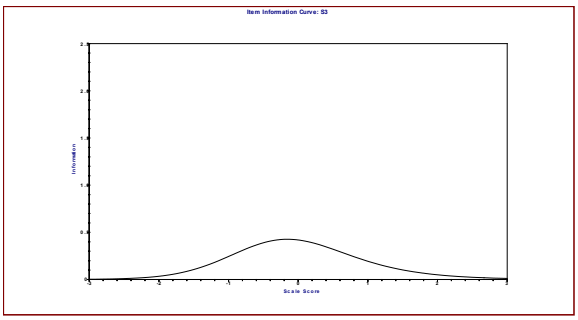
Item 01



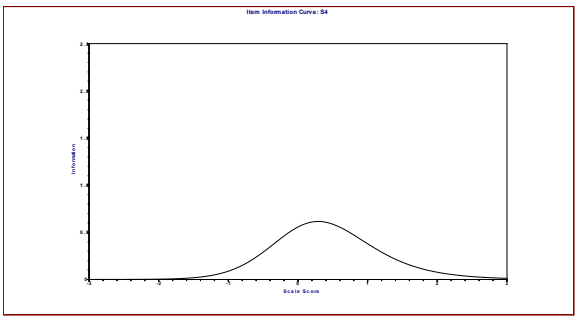
Item 02



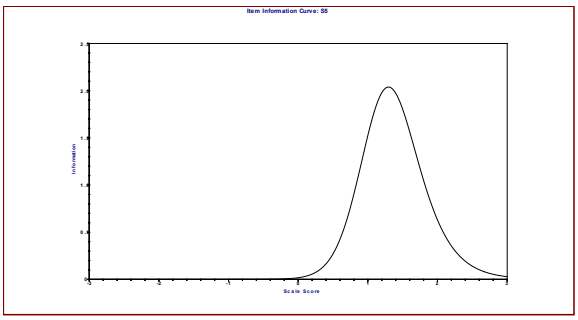
Item 03



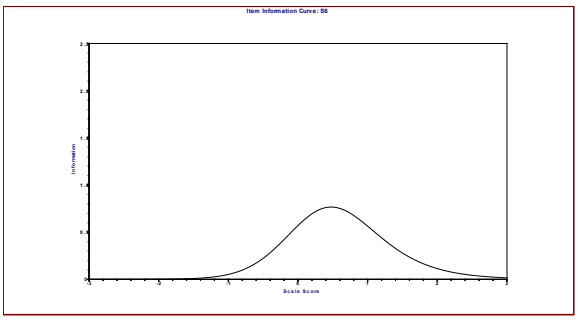
Item 04



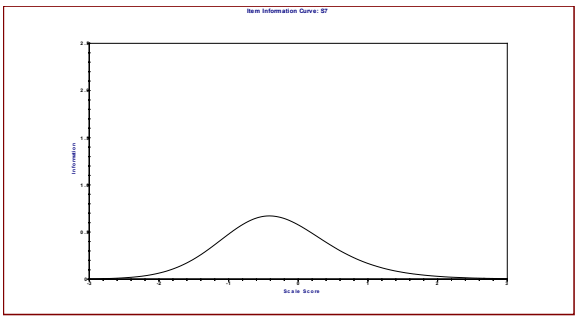
Item 05



Item 06



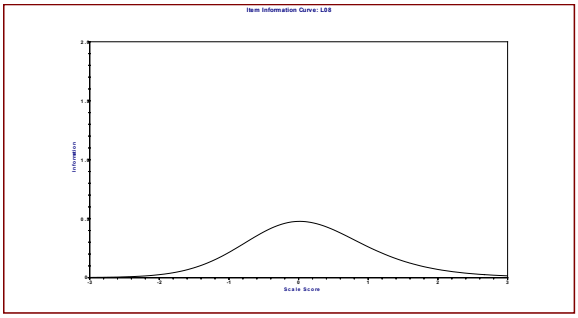
Item 07



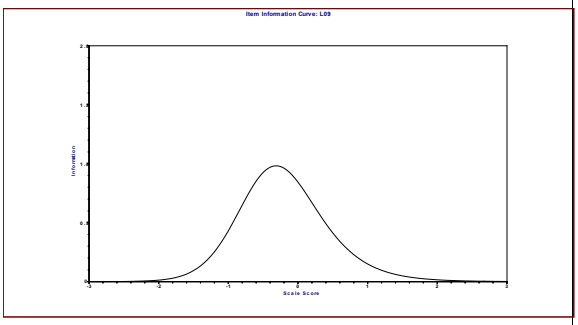
Item 08

Appendix H. Item Information Functions for Listening subdomain in the 2010 GECAT

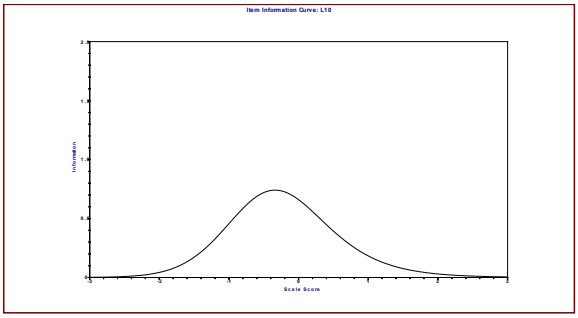
Item 08



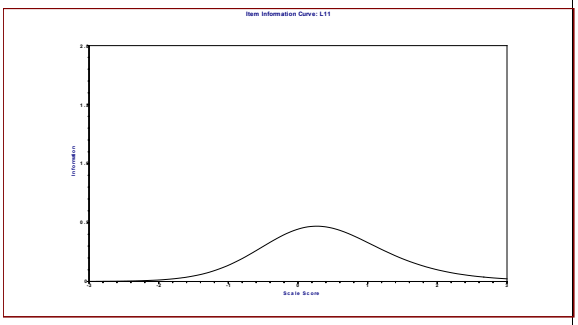
Item 9



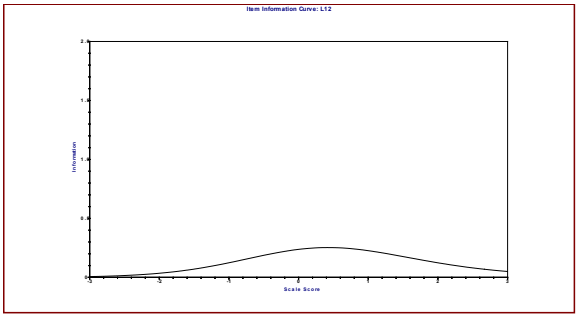
Item 10



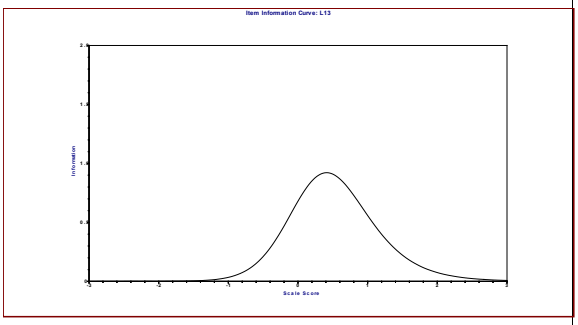
Item 11



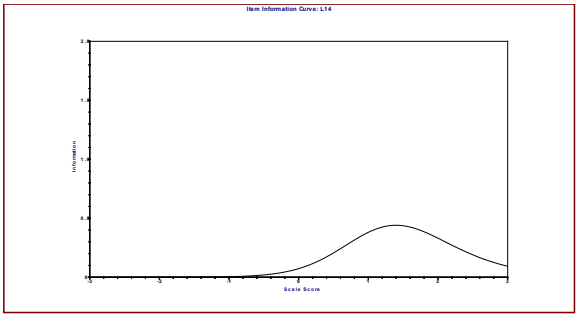
Item 12



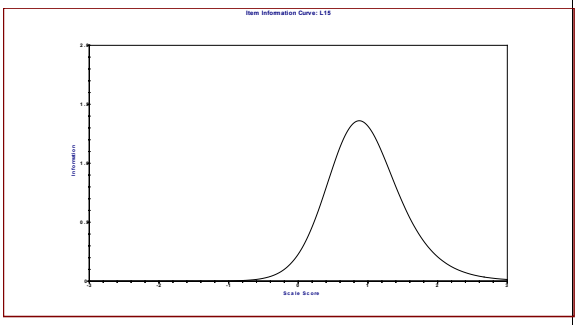
Item 13



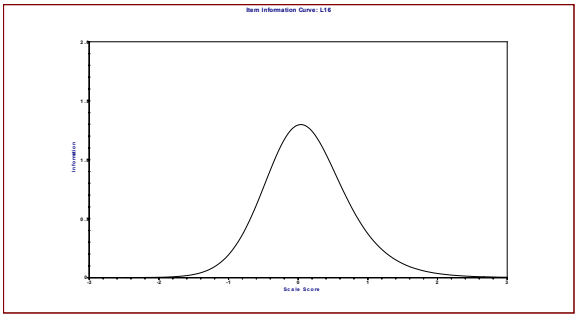
Item 14



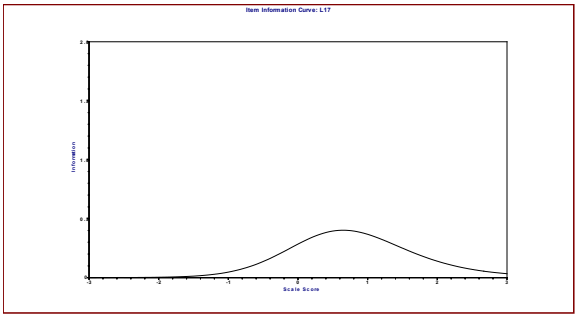
Item 15



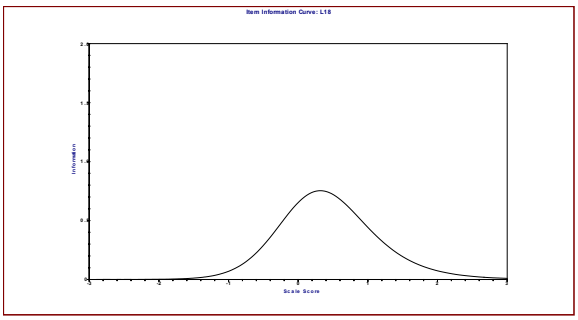
Item 16



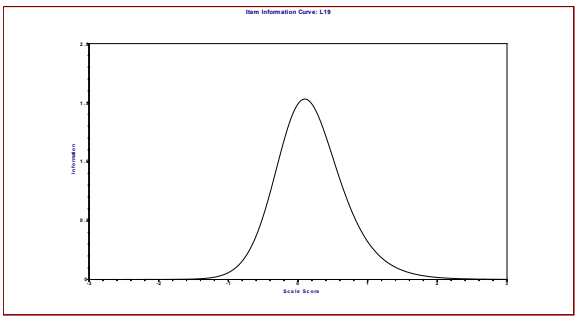
Item 17



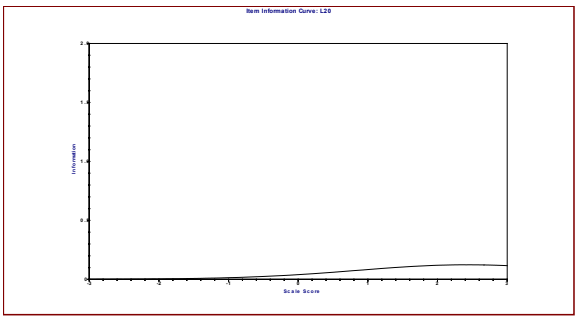
Item 18



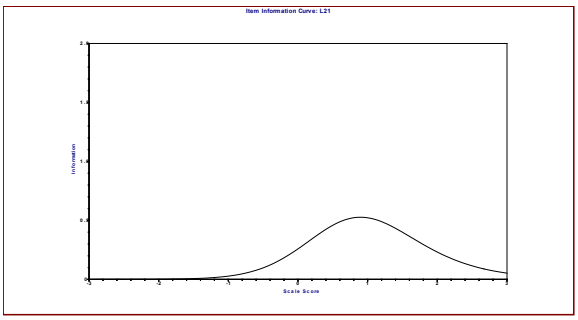
Item 19



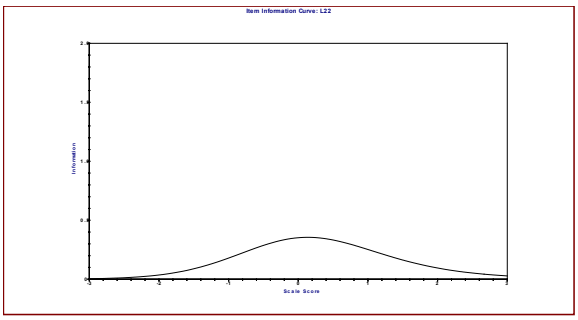
Item 20



Item 21

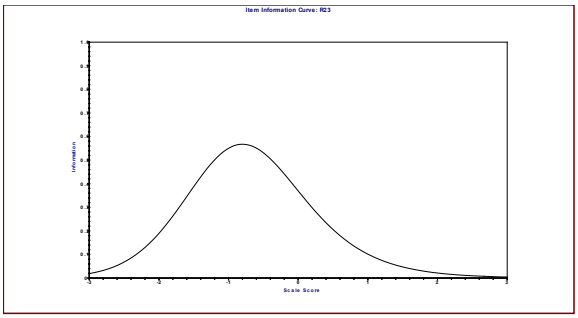


Item 22

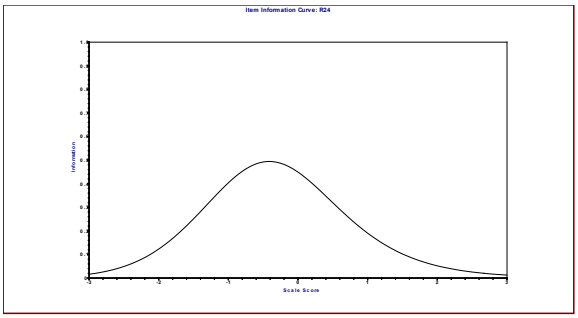


Appendix I. Item Information Functions for the Reading subdomain in the 2010 GECAT

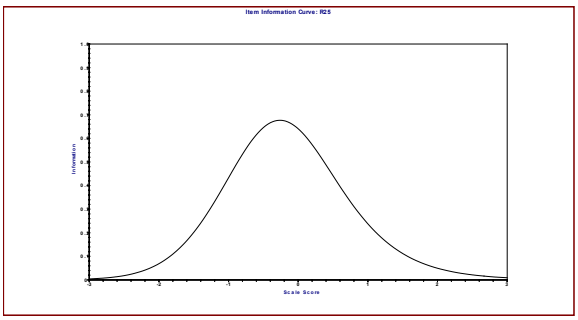
Item 23



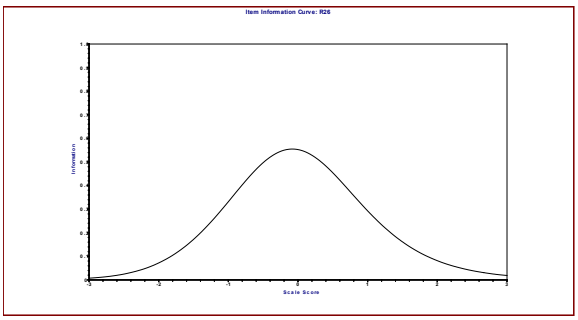
Item 24



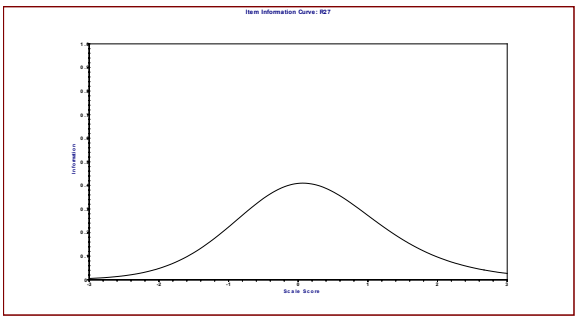
Item 25



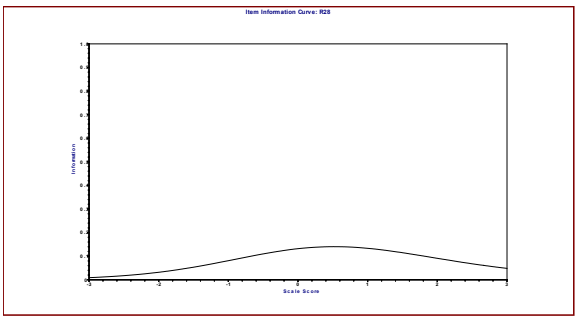
Item 26



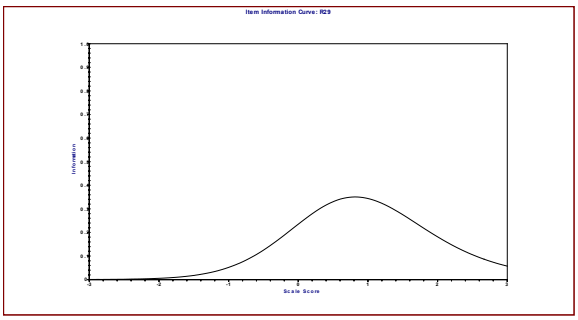
Item 27



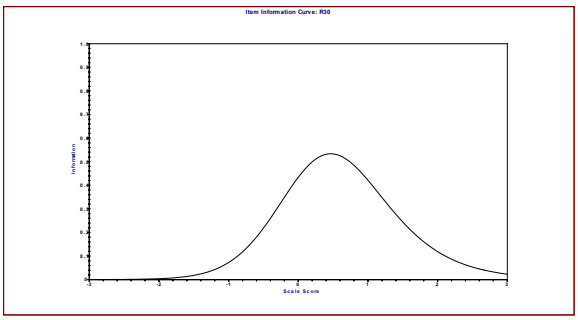
Item 28



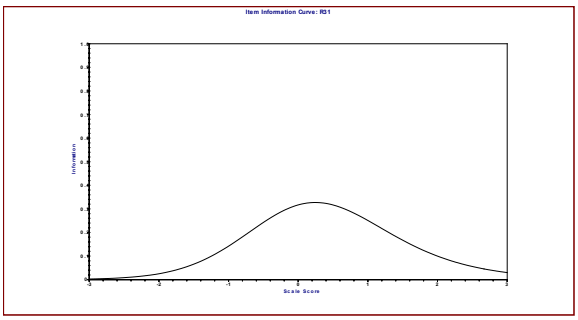
Item 29



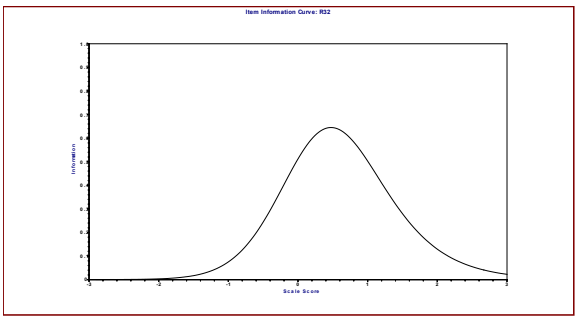
Item 30



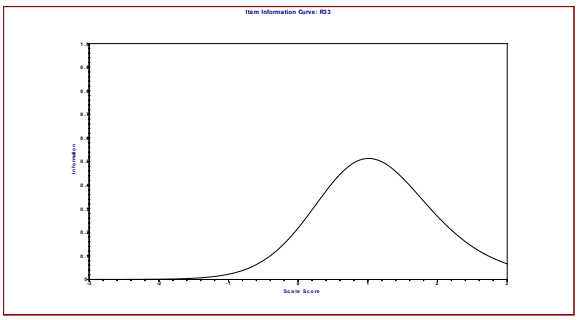
Item 31



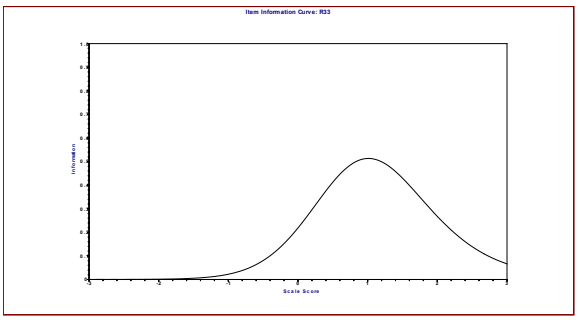
Item 32



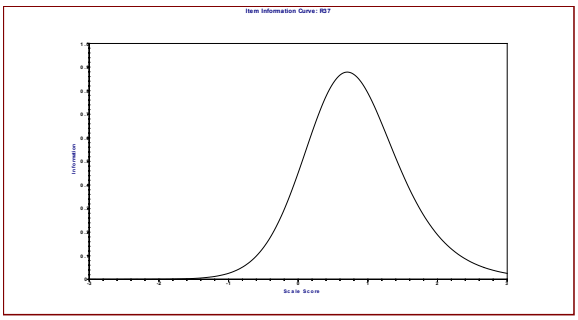
Item 33



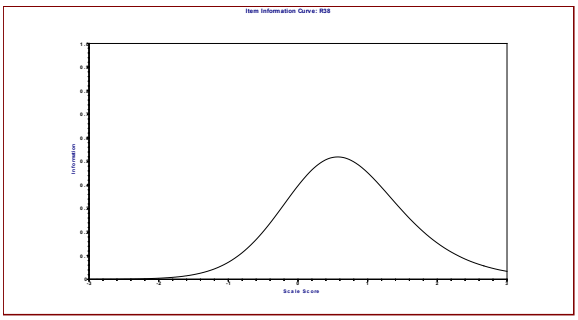
Item 36



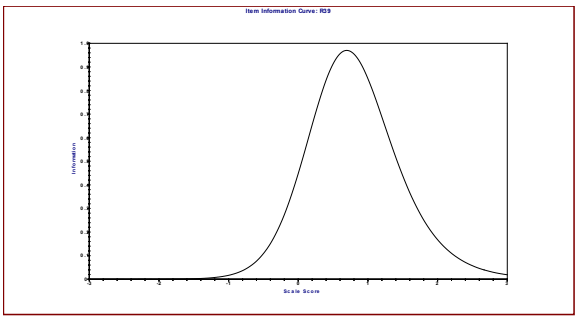
Item 37



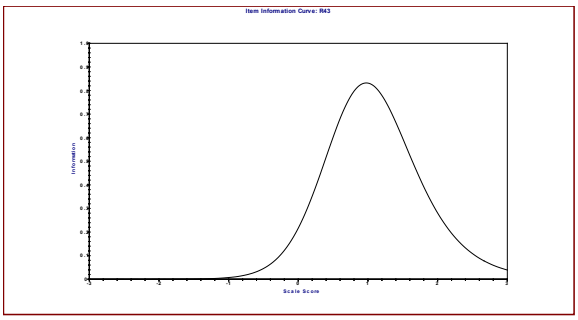
Item 38



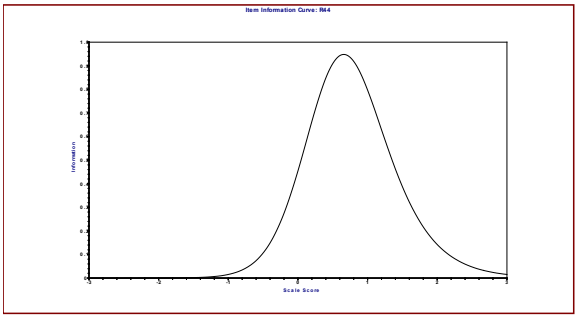
Item 39



Item 43

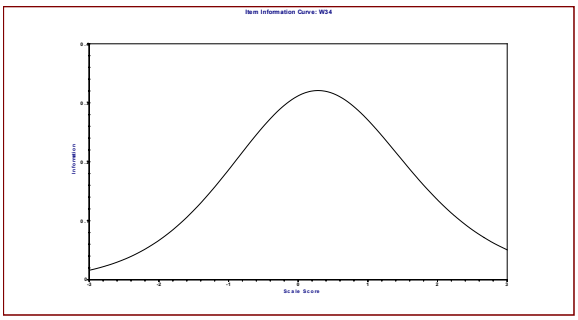


Item 44

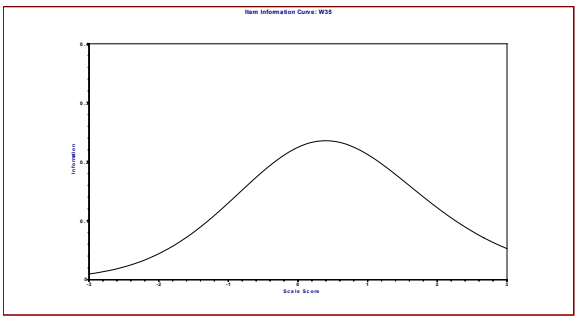


Appendix J. Item Information Functions for the Writing subdomain in the 2010 GECAT

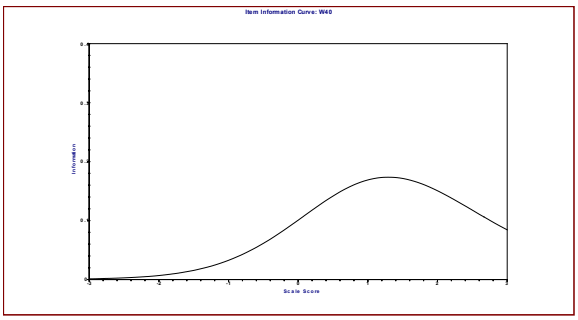
Item 34



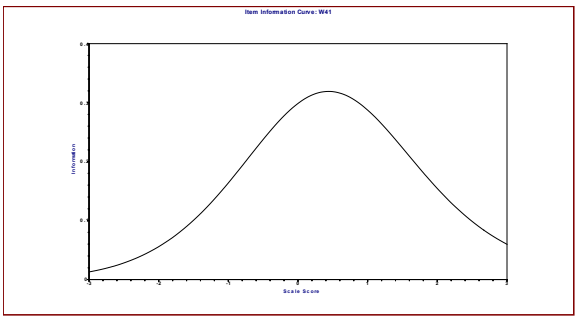
Item 35



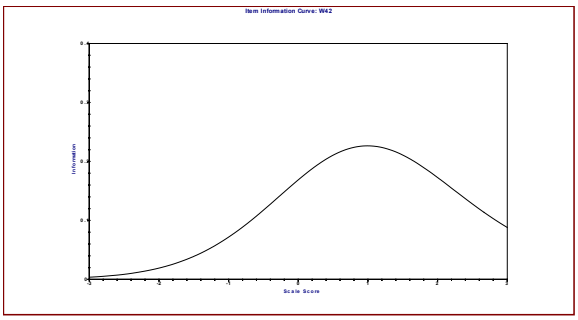
Item 40



Item 41



Item 42

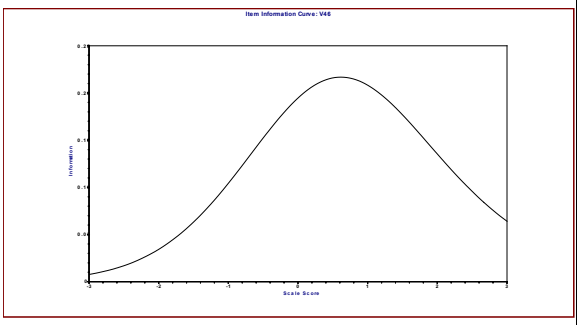


Appendix K. Item Information Functions for the Vocabulary subdomain in the 2010 GECAT

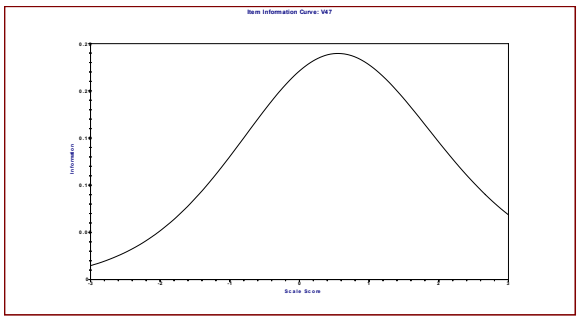
Item 45

Not Estimable

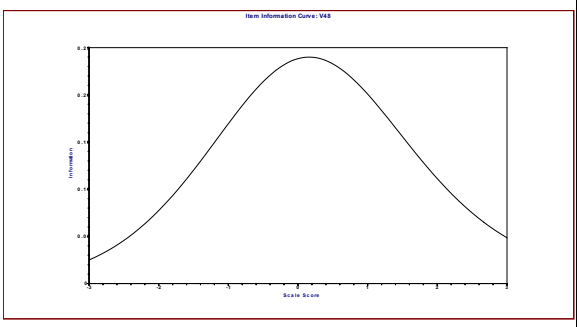
Item 46



Item 47



Item 48



Appendix L. Item Information Functions for the Grammar subdomain in the 2010 GECAT

Item 49	Item 50
Not Estimable	*Not Estimable*

Appendix M. Commands of BILOG-MG for analyzing the items of 2009 GECAT

BILOG-MG COMMANDS for Speaking Subdomain

```
3PL Analysis of 2009 GECAT
SPEAKING SUBTEST ONLY
>GLOBALD FName = 'C:\Users\johnyyun\Desktop\GECAT09.prn',
          NPArm = 3,
          LOGistic;
>LENGTH   NITems = (18);
>INPUT    NTOtal = 18,
          NALT = 5,
          NIDchar = 13;
>ITEMS   INAmes = (S01(1)S18);
>TEST1   TNAme = 'SPEAKING',
          INUmber = (1(1)18);
(13A1, 18A1)
>CALIB ACCel = 1.0000;
```

BILOG-MG COMMANDS for Listenig Subdomain

```
3PL Analysis of 2009 GECAT
LISTENING SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT09.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (14);
>INPUT NTOtal = 14,
       NIDchar = 13;
>ITEMS INAmes = (L19(1)L32);
>TEST1 TNAme = 'LISTNING',
       INUmber = (1(1)14);
(13A1, 18X,14A1)
>CALIB ACCel = 1.0000;
```

BILOG-MG COMMANDS for Reading Subdomain

```
3-PL Analysis of 2009 Korean GECAT
   READING SUBTEST ONLY with SSIGMA PRIOR SPECIFIED
>GLOBAL    DFNAME = 'C:\Users\johnyyun\Desktop\GECAT09.prn',
           NPArm  = 3,
           NTEST = 1,
           LOGistic;
>LENGTH    NITems = (28);
>INPUT     NTOtal =  28,
           NALt   =  10,
           NFMt  = 1,
           NIDchar=  13;
>ITEMS     INUMBERS = (1(1)28),
           INAME = (R49(1)R76);
>TEST1     TNAME = 'READING', INUMBERS = (1(1)28);
```

```
(13A1,48X,28A1)
>CALIB     READPRIOR;
>PRIOR     SSIGMA = (1.05(0)28);
```

BILOG-MG COMMANDS for Writing Subdomain

```
3PL Analysis of 2009 GECAT
WRITING SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT09.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (4);
>INPUT NTOtal = 4,
       NALt = 10,
       NIDchar = 13;
>ITEMS INAmes = (W64,W65,W66,W80);
>TEST1 TNAme = 'WRITING',
       INUmber = (1(1)4);
(13A1, 76X, 4A1)
>CALIB READPRIOR;
>PRIOR SSIGMA = (1.05(0)4);
```

BILOG-MG COMMANDS for Vocabulary Subdomain

```
3PL Analysis of 2009 GECAT
VOCABULARY SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT09.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (7);
>INPUT NTOtal = 7,
       NALt = 10,
       NIDchar = 13;
>ITEMS INAmes = (V33(1)V34, V37(1)V41);
>TEST1 TNAme = 'VOCAB',
       INUmber = (1(1)7);
(13A1, 32X, 2A1, 2X, 5A1)
>CALIB READPRIOR;
>PRIOR SSIGMA = (1.05(0)7);
```

BILOG-MG COMMANDS for Grammar Subdomain

```
3PL Analysis of 2009 GECAT
GRAMMAR SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT09.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (7);
>INPUT NTOtal = 7,
       NALt = 10,
       NIDchar = 13;
```

```
>ITEMS INAmes = (G42(1)G48);
>TEST1 TNAme = 'GRAM',
       INUmber = (1(1)7);
(13A1, 41X, 7A1)
>CALIB READPRIOR;
>PRIOR SSIGMA = (1.05(0)7);
```

Appendix N. Commands of BILOG-MG for analyzing the items of 2010 GECAT

BILOG-MG COMMANDS for Speaking Subdomain

```
3PL Analysis of 2010 GECAT
   SPEAKING SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT10.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (7);
>INPUT  NTOtal = 7,
        NALT = 10,
        NIDchar = 13;
>ITEMS  INAmes = (S01(1)S07);
>TEST1  TNAme = 'SPEAKING',
        INUmber = (1(1)7);
(13A1, 7A1)
>CALIB READPRIOR;
>PRIOR SSIGMA = (1.75(0)7);
```

BILOG-MG COMMANDS for Listeng Subdomain

```
3PL Analysis of 2009 GECAT
LISTENING SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT09.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (14);
>INPUT  NTOtal = 14,
        NIDchar = 13;
>ITEMS  INAmes = (L19(1)L32);
>TEST1  TNAme = 'LISTNING',
        INUmber = (1(1)14);
(13A1, 18X,14A1)
>CALIB ACCel = 1.0000;
```

BILOG-MG COMMANDS for Reading Subdomain

```
3PL Analysis of 2010 GECAT
READING SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT10.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (17);
>INPUT  NTOtal = 17,
        NALt = 10,
        NIDchar = 13;
>ITEMS  INAmes = (R23(1)R33,R36(1)R39,R43(1)R44);
>TEST1  TNAme = 'READING',
        INUmber = (1(1)17);
(13A1, 22X, 17A1)
>CALIB READPRIOR;
>PRIOR SSIGMA = (1.15(0)17);
```

BILOG-MG COMMANDS for Writing Subdomain

```
3PL Analysis of 2010 GECAT
WRITING SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT10.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (5);
>INPUT  NTOtal = 5,
        NALt = 10,
        NIDchar = 13;
>ITEMS  INAmes = (W34(1)W35,W40(1)W42);
>TEST1  TNAme = 'WRITING',
        INUmber = (1(1)5);
(13A1, 39X, 5A1)
>CALIB READPRIOR;
>PRIOR SSIGMA = (1.05(0)5);
```

BILOG-MG COMMANDS for Vocabulary Subdomain

```
3PL Analysis of 2010 GECAT
VOCABULARY SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT10.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (3);
>INPUT  NTOtal = 3,
        NALt = 10,
        NIDchar = 13;
>ITEMS  INAmes = (V46(1)V48);
>TEST1  TNAme = 'VACAB',
        INUmber = (1(1)3);
(13A1, 45X, 4A1)
>CALIB READPRIOR;
>PRIOR SSIGMA = (1.01(0)3);
```

BILOG-MG COMMANDS for Grammar Subdomain

```
3PL Analysis of 2010 GECAT
GRAMMAR SUBTEST ONLY
>GLOBAL DFName = 'C:\Users\johnyyun\Desktop\GECAT10.prn',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (2);
>INPUT  NTOtal = 2,
        NALt = 10,
        NIDchar = 13;
>ITEMS  INAmes = (G49(1)G50);
>TEST1  TNAme = 'GRAMM',
        INUmber = (1(1)2);
```

```
(13A1, 48X, 2A1)
>CALIB READPRIOR;
>PRIOR SSIGMA = (1.01(0)2);
```