# TOEFL IN A BRIEF HISTORICAL OVERVIEW FROM PBT TO IBT

Gunadi H. Sulistyo

Jurusan Sastra Inggris Fak. Sastra Universitas Negeri Malang

**Abstract**: This article briefly reviews the development of TOEFL as a widely acknowledged version of English proficiency test for non-native users. The specific aspects on review are the nature of TOEFL as a testing instrument, its historical development from the perspective of language and test development theories, and the testing formats of the language aspects in both earlier and later versions of TOEFL. Also on elaboration is the scoring comparison applied to both versions..

**Kata kunci**: TOEFL, development, version

The Test of English as a Foreign Language (henceforth TOEFL) has enjoyed its prestigious status as a standardized test widely used across nations of more than one hundred countries since its initial establishment in early 1960s. It has been utilized as a means of measuring the proficiency of non-native speakers of English, as its name demonstrates, in English as a foreign language, in particular for academic purposes. Not only international educational institutions, several domestic higher-learning institutions as well as non educa-tional agencies have also made use of the score of individuals taking TOEFL as a requirement of not only admission, recruit-ment, but also exit purposes. This implies that many have relied on TOEFL as a dependable tool that can provide good evidence of one’s proficiency in English as a foreign language.

It is believed that interest in taking TOEFL has been increasing as shown in the number of the prospective TOEFL takers from year to year. This indicates that the needs of TOEFL scores have also boosted from year to year. As a consequence, the needs of TOEFL training are also inevitable although it is not clear whether those prospective TOEFL takers take the test for their further studies abroad or for any other purposes. What is obvious then is that such a demanding context triggers the establishment of preparatory courses that burgeon ubiquitously. It is an undeniable fact that such preparatory courses in a way play a role in catering for the needs of the prospective TOEFL candidates of the TOEFL scores a part from any interest in their establishment.

While the technologies of testing adopted by TOEFL have advanced more rapidly, and at the same time, while the need of the TOEFL scores tends to be increasing in number, it seems that those preparatory courses have not been able completely to catch up with, in particular, the advances of testing technologies employed by TOEFL. For example, on one side

TOEFL currently has adopted the socalled next generation version of the internet based testing practices, or iBT (henceforth) since 2005, which has been a significant shift from older TOEFL versions of computer-based TOEFL (cBT for short) as well as paper-and-pencil based TOEFL (pBT, henceforth). On the other side, those training courses mostly still deal with the pBT version. Therefore, there seems to be a need of those running preparatory TOEFL programs to seek for a more wide-ranging picture of all the practices that TOEFL has undergone so far. There are several reasons for this necessity. In the first place, those courses will have accurate visions in provi-ding the prospective TOEFL candidates with accurate information concerning the type of language skills these candidates pursue. Next, those courses in effect will try their best to provide the prospective TOEFL candidates with appropriate language skills that reflect most closely real academic life. Also, those courses will always try to update themselves with TOEFL's most current technologies and practices, which ultimately will benefit the prospective TOEFL candidates from joining the courses they offer.

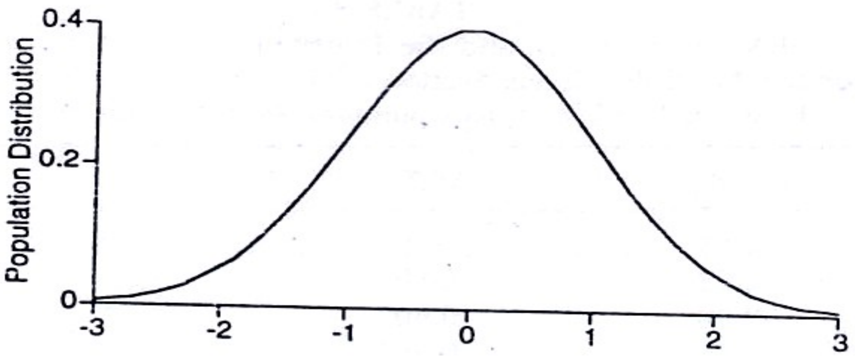
This piece of paper is aimed briefly at reviewing TOEFL from its first version, pBT, to its current version, iBT. For the purpose, several topics will be dealt with, covering first the discussion that touches more on the conceptual ground of the nature of TOEFL. The next part will deal with the development of TOEFL. Following this part is the presentation of the components that make up each existing TOEFL. Scoring matters constitute the next part. Finally, the last part concludes the paper.

.

## NATURE

By purpose, TOEFL can be categorized as a proficiency test. Brown (2005:8) defines a proficiency test as the test that has the function to '…assess the general knowledge or skills commonly required or prerequisite to entry into (or exemption from) a group of similar institutions.' The generality level of proficiency of the test implies first that TOEFL is not linkable to a particular school syllabus or curriculum because TOEFL is established on the basis of concept of general language ability. Simply, in a more operational term the materials that are contained in TOEFL do not reflect the instructional materials of a particular syllabus or curriculum.

Secondly, as a test established on general language ability, TOEFL is necessarily a norm-referenced test. This kind of test is to produce scores that can spread individuals taking the test along the ability line ranging from the least able to the most able. Psychometrically, such a test needs to be able to put an individual in a point along the ability line ranging from - ∞ to + ∞. This view also posits that the number of the people with the ability close to these two extreme points ( - ∞ and + ∞) is fewer than that of the people with the ability around the average. This is what is commonly known as assumption of normality. When plotted, each level of ability in the universe of a particular group necessarily forms a bell, and its distribution is commonly known as a bell-shaped distribution



(Rosa, et al., 2001:261)

Figure 1: Assumed Normal Distribution of Ability

In this view one's performance in TOEFL is to be compared to another's performance in the same test. Consider the following figure.

A                                                                    B

-∞      _______.____________________________________________.__________      + ∞

Figure 2: A Hypothetical Standing Ability of Two Individuals in Ability Line

In the figure displayed above the standing of A is relative to the standing of B in an ability line that ranges from - ∞ to + ∞ with B being considered more able than A in the line. This follows then that, the generality nature of a proficiency test makes it possible for a comparison of not just individual but also groups.

Also, as Brown (2005:9) puts it to say as a test of language proficiency, TOEFL can play a role as an external measure which is neutral to individuals as well as groups. In the case of academic contexts within English-speaking countries, say the USA or Canada, an individual's obtained TOEFL score will make it possible to determine whether an individual fits in a particular program or not. Similarly, an individual's score upon completion of a TOEFL program will be able to be used as an indicator of his/her proficiency level that can predict his/her success in other context.

## DEVELOPMENT OF TOEFL: HISTORICAL PERSPECTIVES

Thus far, TOEFL has witnessed three successive major formats: pBT, cBT, and iBT. Initiated by an American council on the testing of English as a foreign language in the early of 1962, TOEFL and its historical development can be viewed from two angles. The first perspective is concerned with development of TOEFL as seen from the underlying concept of language ability; and the second angle deals with advances in the testing technology that characterize TOEFL.

**The pBT version**. Viewed from the linguistic perspective, TOEFL originally adopts the structural linguistic view and this is obvious in the pBT format. This structural linguistic view believes that language is divisible in nature. Just recall the concept of duality in language. Language is of two main layers: the layer of form and that of meaning. The former is concrete; the latter abstract. As such the former is believed to be more learnable than the latter. The layer of form consists of other divisible layers, from phonemes as the smallest unit to syntactic constructions as the largest. Language ability is conceptualized as the mastery of the layers one by one. This follows that there is a need to test one's mastery of these layers bit by bit. In TOEFL this view is reflected clearly in the importance of having accuracy of grammatical form or testing a TOEFL taker's grammatical knowledge.

Language is also conceived to comprise interacting two components: language skills and language components. Language skills refer to the modes through which language components may be realized, which include listening, speaking, reading, and writing. Language components include grammar/structure, vocabulary, phonology/orthography and fluency. Harris (1969:11) neatly illustrates the relation between language skills and language components as the following matrix suggests.

| Language Skills | Language Components | | | |
|---|---|---|---|---|
| | Grammar / Structure | Vocabulary | Phonology/ Orthography | Rate and General Fluency |
| Listening | v | v | v | v |
| Speaking | v | v | v | v |
| Reading | v | v | v | v |
| Writing | v | v | v | v |

(adapted from Harris, 1969:11)

Figure 3: Matrix on Language Skills and Components

In the earlier format of TOEFL, the adoption of the structural linguistic view is obvious. For instance, in the pBT the test is comprised of three sub tests as Listening Section, Grammar and Written Expression Section, and Reading Comprehension Section. The other two, which are normally tested separately, include the Test of Written English (TWE) and the Test of Spoken English (TSE).

The influence of the structural linguistic view, when examined further, is also obvious in the formulation of test items. For instance, the listening section is composed of three independent parts reflecting the existence of linguistic layers: comprehension of fragmented sentences, comprehension of dialogs, and comprehension of texts larger than dialogs/monologs. In the comprehension of fragmented sentences, accuracy in frequently grammatical points is tested.

In addition to these, the second section, Grammar and Written Expression Section, clearly reflects the structural linguistic view. In this part, a sentence of particular grammatical complexities is presented with one part containing a grammatical mistake. The lexical meaning of the sentence is as far as possible kept plausible. The TOEFL takers are to identify the mistake. Very frequently the mistake is of 'local errors', which do not potentially interfere with communication. This section has a typical item as follows

The Peace Corps was establish on March 1, 1961 by then President John F. Kennedy.
A B C D

In such an item, rather than communicativeness of an expression, grammatical sensitivity of the expression is being assessed. This seems to be typical of the structural linguistic view for accuracy of form constitutes an utmost important prerequisite for language mastery.

Reading section also clearly suggests the influence of the structural linguistic view. This section may be introduced with a test on vocabulary items which are presented in sentential contexts. This part also frequently begins with a short text, presumably a paragraph with questions following the text. Items on testing understanding the meaning of a word in context are also included as a question.

Seen from the testing technology adopted in the pBT version, the test format used in TOEFL fits the divisibility nature of language. Apart from its test of written English (TWE) and test of spoken English (TSE), TOEFL employs the multiple-choice type with four options. In this selective type of response, there is a stem that functions as a stimulus to which the TOEFL takers will respond. Following the stimulus are the alternatives with one correct answer and three distracters for the TOEFL takers to select. The use of multiple-choice enables the language elements to be measured bit by bit. In addition, the presentation of items in the test follows an order with an increasing level of difficulty. However, the items presentation is fixed in that the TOEFL takers have no choice in completing the items presented to them whether or not the level difficulty (*p*) of the items fit their language ability. The earlier items are those with medium level of difficulty, or $.60 \leq p \leq .80$ (Crocker and Algina, 1986:312). This feature is understandable due to the nature of the pBT format which does not permit the level of difficulty of the items to vary along the line with the ability level of the test takers.

**The cBT version**. Seen from its underlying linguistic theory, basically the cBT version is still characterized by the structural linguistic views. Thus, there have not been significant changes in the concepts reflecting general language ability adopted in the cBT version. Just like the pBT version, in the cBT earlier version, the test is comprised of three sub tests as Listening Section, Grammar and Written Expression Section, and Reading Comprehension Section. The other two, which are normally tested separately, still include the Test of Written English (TWE) and the Test of Spoken English (TSE). The test format used also remains the multiple-choice type with four selective alternatives. In later years, however, the cBT version adopted more and more communicative views of language competence.

The cBT version, however, may be classified into two in terms of the testing technology adopted in this version. The first cBT type is essentially like the pBT version in that it still assesses listening, grammar and written expression, and reading with the same sub tests distributed in 3 main sections. The test format is the same: the multiple-choice type with four selective alternatives and the items presentation is fixed with an increasing level of difficulty. As such the first cBT type may be known as the non-adaptive type of the cBT version. This first type is basically the pBT version which is made computerized.

In the initial stage, the second cBT type is essentially like the non adaptive cBT one in terms of contents and format: measurement of listening comprehension, sensitivity on grammar and written expressions, and reading comprehension with the same sub tests distributed in 3 main sections; the test format being the same: the multiple-choice type with four selective alternatives. What differs essentially lies in the presentation of test items. This type still adopts a strategy with an increasing level of difficulty with earlier items being those with a medium level of difficulty. However, the items following the first items may vary depending on the response of the TOEFL takers. Figure 4 describes the scheme of adaptive testing.

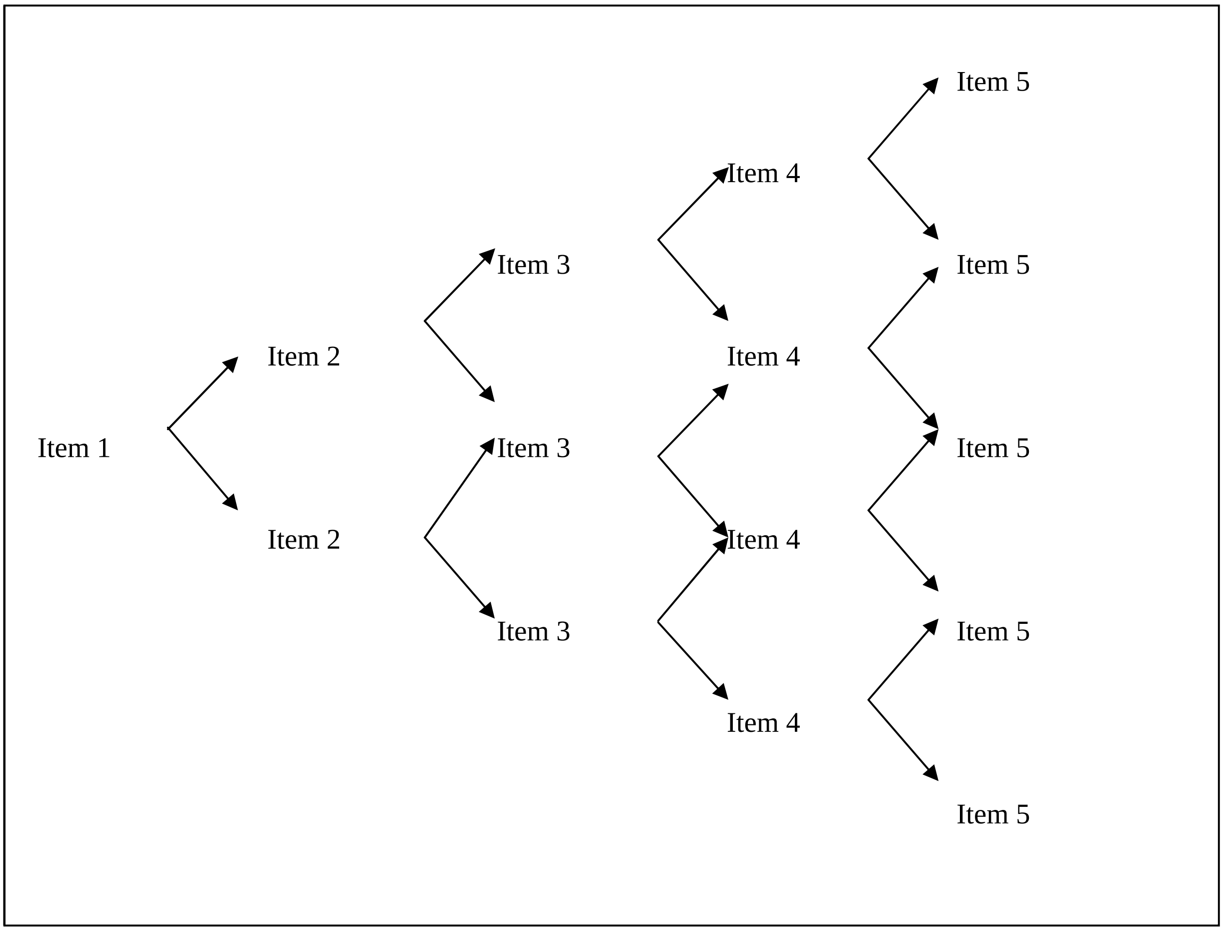


Figure 4: One Simple Scheme of Adaptive Testing

A correct response made by the test taker will be a stimulus for the computer to process a more difficult item than the earlier to be completed next by the test taker. Conversely, a wrong answer will lead to the presentation of an easier item to the test taker to be responded. Thus, the level of difficulty of the test items adapts the test taker's level of ability. Because of this mechanism, the second cBT type is commonly known as the adaptive version of TOEFL.

Unlike the non adaptive one, the adaptive version of TOEFL is based on the work of the modern test theory or the item response theory in which items are made invariant across test takers (Hulin, Drasgow, and Parsons, 1983:43), which means that no matter who responds to the items, the characteristics of the items remain the same. Thus, under this scheme a test taker will only respond to the items that fit his/her level of ability. The adaptive version is further facilitated with advances in computing technology where manual computations will be extremely time-consuming and tiring with no assurance of accuracy.

In addition to these, in the later development, a new cBT begins to include important views in more recent advances in the concept of communicative language use (TOEFL Internet-Based Test, 2007:3). Of the views in the new concept of communicative language use, language is considered more functionally with a focus as a means of communications. While verbal communications mean the realization of competence in performance, the new CBT goes on this tract and is concerned with the inclusion of macro language skills: listening, speaking, reading, and writing. Language components are tested in integration within language skills rather than in isolation.

In addition to this shift, the new cBT also adopts a new testing mode that provides TOEFL takers with more opportunities to demonstrate their command in utilizing macro skills by way of constructing direct responses, thus beginning to leave the selective-response format behind. Another shift taking place in the later development of the new cBT relates with the language processing. Unlike the old cBT version which is essentially the pBT version made computerized and elicits language abilities through a discrete-mode of testing, the newer cBT begins to include test tasks that would process language in a more integrative fashion in terms of language skills. Thus, writing or speaking tasks may have a relation with a reading or listening task. Prior to producing English through writing or speaking tasks, a candidate may be required to incorporate pieces of information that appear in a reading or listening task.

**The iBT version**. This version was launched in 2005, and is gradually expected to substitute the role of both the cBT and the pBT versions. The introduction of the new cBT version plays a critical role to the establishment of the iBT version in that the new cBT version lays a strong transitional bridge on which to step onto the era of the iBT version.

As has been discussed previously, from linguistic standpoints, the later cBT version has endeavored to feature the principles of the communication-movement in language testing in the corresponding TOEFL sub tests. As the further development of the later cBT version, the iBT version shares similar features with the later cBT version. From linguistic standpoints, the iBT version is also characterized by the need to test the candidates' functional language skills. It is designed to measure the integrated use of macro skills: listening, speaking, reading and writing. Also, in this iBT version, academic settings and themes are more emphasized.

Introduced worldwide in the period of 2005 - 2006, the iBT version, as its name indicates, makes functional use of information and communication technology (ICT). The TOEFL tasks in the iBT version are delivered through the internet from Educational Testing Services (ETS) to the authorized testing centers where the candidates are pooled to complete the tasks.

## BUILDING BLOCKS INSIDE ALL TOEFL VERSIONS

All the three TOEFL versions: pBT, cBT, and iBT essentially have their own characteristics viewed from two main points: what to be tested and how to test it. What follows is a brief account of each of the TOEFL versions seen from these two points.

In terms of what to test, the pBT version, as it was influenced by the structural linguistic views, is characterized by discrete-testing practices. One (or maybe two) language component is obviously tested, namely grammar, under a separate section. Vocabulary is also tested, but is commonly put under reading. Two macro skills are tested, namely listening and reading. These three aspects: grammar, listening, and reading all together comprise one battery commonly known as the paper-based TOEFL. Two other macro skills speaking and writing are also tested as independent sets known as the Test of Spoken English (TSE) and the Test of Written English (TWE) respectively.

The listening section of the pBT version normally consists of a variety tasks assessing three or four aspects: sentence level comprehension, comprehension of dialogs, comprehension of extended conversations, and comprehension of mini talks. These aspects clearly indicate levels of how language is believed to be constructed. The social themes presented to assess these aspects are commonly of general interest. In mini talks, however, mini lectures are also presented. This is intended to represent academic settings.

The grammar section mainly aims at testing grammatical accuracy and, in one sense, grammar sensitivity. The grammatical points to be tested include a variety of English grammar aspects such as verbs, auxiliary verbs, nouns, pronouns, modifiers, comparatives, connectors, sentences and clauses, relationship of ideas, agreement, introductory verbal modifiers, parallel structures, redundancy, and word choice (Sharpe, 2005:86-113).

The reading section may consist of two main aspects to be tested: vocabulary and reading. Vocabulary includes the testing of word meanings and/or meanings of words in sentential contexts (Jenskins-Murphy, 1981). This includes among other things testing of shades of meaning of words, synonym, antonym, word-part clues, denotation, and connotation. Reading aims at assessing various micro reading skills, like understanding main idea, understanding supporting ideas/details, understanding organization of the text, understanding implied details, understanding word meaning, understanding pronoun reference, and understanding the writer's tone of writing (Phillips, 1989). The rhetoric modes of the text include among other things narration, definition/illustration, classification, comparison, contrast, cause, effect, persuasion/justification, problem/solution (Sharpe, 2005:122-251).

The writing test assesses a candidate's ability in writing a piece of essay of expository or persuasive modes. Focus of testing is placed on the candidate's ability to organize ideas using accurate grammar, vocabulary, spelling and mechanics. Just like the writing test, the speaking test aims at assessing a candidate's clarity in expressing his/her ideas in spoken English of exposition. The candidate's use of grammar, vocabulary, and pronunciation as well as idea organizations are also evaluated.

In terms of how to test, the pBT version is basically non adaptive testing. The items are arranged with their fixed yet increasing level of difficulty along the items in the corresponding battery. Aspects to be tested are organized into sections indicating particular abilities to be assessed. For instance, the listening comprehension section is organized into three main parts: sentence comprehension or dialog part of about 30 questions, extended conversation part of about 5 questions, and mini-talk part of about 5 questions. The grammar and written expression section is differentiated into two parts: error recognition and completion with about 20 questions each. The reading comprehension section may take the form of 4-5 reading passages with about 7-9 comprehension questions following them. Vocabulary items are included in this part. Writing is an independent task in the pBt version which requires the test takers to respond to a writing task using their hand writing. Normally, it takes 3 hours to accomplish the pBT version.

Basically there are not many differences between the pBT and cBT versions in terms of what to test. Thus, the cBT version tests both language components (grammar and possibly vocabulary) and skills (listening, reading, and writing). Slight differences are observed, however. In the listening section of the cBT section, less dialogs (about 20) are presented with one question each. Extended conversations in the pBT versions are modified a bit into short conversations in the cBT version with about the same number i.e. 3 short conversations with about 3 questions each. The pBT's mini-talks are also modified and specified as mini-lectures and discussions in the cBT versions with about 5 questions each. Also, the reading section of the pBT and cBT versions is almost similar in terms of the number of passages and comprehension questions following the passages, figuring at about 5 passages and 10 questions each respectively. In terms of the writing tasks, the pBT and cBT versions share similarities i.e. one independent task, focusing on exposition/persuasion. The scoring of writing, however, is combined with the structure section. Finally, the number of items in the structure section is reduced in the cBT version to about 25 questions from about 40 questions in the pBT version. All the sub tests in the cBT version need 3.5 hours to accomplish.

In terms of how to test, as has been dealt with previously, the cBT version is psychometrically speaking adaptive, in particular in the listening and the structure sessions, but not for the reading and writing sections. Initial numbers are presented with items of moderate levels of difficulty. The presentation of other items following these initial numbers depends on the answer to these initial numbers. An item with a more difficult level follows a correct answer; an item with less difficult level tags along with an incorrect answer, and so on. Simply, the presentation of an item adapts the TOEFL taker's level of ability or technically known as *theta* and symbolized as *T* (Hulin, Dragow, and Parsons, 1983:26).

The iBT version as the next generation TOEFL may be considered as a significant innovation in the construction of TOEFL. It comes up utterly with not only a new format along with all macro skills but also a new presentation. As aforementioned, the iBT version captures the views of the communicative approach. These are clearly reflected in the inclusion of all macro language skills: listening, speaking, reading, and writing as integrated sub tests while excluding language components: grammar and vocabulary as a discrete sub test. In addition, academic themes are also considered in the new version. This is to say that integration of language components into language skills, and integrated processing of information within language skills as well as an emphasis on academic settings has characterized the iBT version strongly.

A further examination on each sub section of the iBT version reveals substantial changes that have been made. In the listening section, more genuine academic conversations and lectures are presented. Other aspects include questions that probe further about the speaker's mood, feeling, purpose and drive are also posed to the test takers. More naturally, note taking, which did not appear in the previous versions, is permitted in the iBT version. In terms of types of aural stimulus, listening tasks include two main formats: lectures (6 texts) as well as classroom discussion with corresponding questions (about 5) each, and academic setting conversations (about 3) accompanied with 5 questions each. With this format, the iBT version clearly presents more contextual academic materials and eliminates dialogs of general themes that are normally presented as fragments. The micro listening skills assessed include identification of main ideas, supporting details, inferences, functions, and organizational structure of the text.

Unlike in the previous versions, where speaking was an optional sub test tested only on particular testing dates, in the iBT version speaking is an integral part of the battery. In all there are 6 tasks comprising of two independent tasks dealing with expression of an opinion on a known topic of academic matters and four integrated tasks requiring speaking on the basis of information picked up in the listening and reading sections. More specifically, in the integrated mode involving reading, listening, and speaking, test takers are first to read a text, listen to a text, and then to speak the relationship of ideas in the two texts. While in the integration of listening and speaking, the test takers are first to listen long texts and then to make a summary and express, or defend their opinions on the information contained in the texts with clarity, coherence and accuracy.

Reading tasks in the iBT version also take a different format. In the new format 3-5 long texts of academic themes are presented, followed with 12-14 questions each text, covering the micro reading skills about main ideas, supporting details, inferences, restatements, sentence insertion, language functions, organization of ideas. Rather than recognize and make a choice as has happened in the old versions, in the iBT version the test takers are to make responses in the form of categorizing information, filling out tables or charts, making or completing summaries, or paraphrasing.

Just like speaking, writing is also integral to the whole TOEFL battery in the iBT version. There are two tasks: one integrated task which requires the test takers to write the relationship of ideas of the academic texts they have read and heard; one independent task which requires the test takers to support a personal opinion in the form of an essay.

Advances in the test construction of the iBT version have not, however, happened in the area of presenting adaptive items. Thus, in accomplishing items of the iBT version, test takers are presented with the same array of test problems.

The presentation described above clearly indicates that the iBT version of TOEFL differs markedly from its other two predecessors. The differences are obvious from two aspects: what to test and how to test, but not in terms of adaptability of the test tasks to the test takers' level of ability.

## SCORES IN COMPARISON

The changes in the format of all TOEFL versions are also accompanied with changes in the scale used to score the TOEFL takers. These changes are necessary. There are reasons for the changes. In the first place the components to be tested change. As a result, there is a need to accommodate the scoring system in evaluating the TOEFL takers' score in each of the components to be tested. Secondly, the underlying philosophies of the versions have also shifted from structural to communicative, thus responding to more recent advances to the theory of language. Most importantly, a more meaningful interpretation is needed as what proficiency is indicated by the score on taking TOEFL of different versions. It is reported that extensive studies on the scoring comparison have been conducted involving a number of 3,000 from 30 countries between the period of 2003-2004 (ETS, undated:4)

Score comparison in all TOEFL versions may be viewed from the total score or from each corresponding separate sub test: listening, reading and writing. Meanwhile, scores obtainable from the grammar section are relevant to be compared because they are only available from pBT and cBT versions. However, for a more meaningful interpretation of the score obtained, separate scores are more desirable because these separate scores are more informative than the total score. The following table presents TOEFL score scales of the three versions.

Table 1: TOEFL Score Scale Comparison

| TOEFL Version | Aspects to Be Tested | | | | | Total Score |
|---|---|---|---|---|---|---|
| | Listening | Structure | Speaking | Reading | Writing | |
| iBT | 0 - 30 | n/a | 0 - 30 | 0 - 30 | 0 - 30 | 120 |
| cBT | 0 – 30 | 0 - 30 | n/a | 0 - 30 | combined with Structure | 300 |
| pBT | 31 - 68 | 31 - 68 | n/a | 31 - 67 | n/a | 677 |

(adapted from ETS, undated:5)

The table clearly shows that the total score for each version differs markedly with the iBT total score of 120, cBT of 300 and pBT of 677. These total scores also imply a score transformation in particular with cBT and pBT, to reach a total score. It is true that in cBT and more obviously pBT, score transformation is performed. In order to estimate a score on a sub test, a table of score conversion is required (Sulistyo, 2001). This explains that totaling a maximum score in each sub test in cBT and pBT does not automatically yields a total score of each of these versions. This is not the case with the iBT, where totaling the maximum score of each sub test automatically yields the total score of the version. Thus there seems to be a more simplification in the scoring scale in the more recent version.

A closer look at each of sub test in each version also reveals how scoring in each version is performed in a different way. With cBT and iBT, the minimum total score in each aspect to be tested is 0 (zero) and the maximum score is 30. Thus in terms of the assigning the lowest and the highest scores they share the same ground. The yielding of the total score, however, is different. pBT obviously utilizes a different score assignment, the lowest score being 31 while the highest 67 or 68 points.

It has been touched upon previously that pBT and iBt do not utilize an adaptive mode of testing unlike the cBT version. However, what is unclear in relation to the scoring mechanism, particularly with the cBT version is whether a transformation from raw scores to ability scores is performed is carried out or not such as that normally adopted in the application of the item response theory in real testing context (Baker, 1985). In the traditional mode of score interpretations, raw scores are assumed to reflect abilities. This is unlike the practice in the modern mode of score interpretations where abilities are not just the sum of the correct answers (Lord, 1980). One’s ability or known as *theta* (*T*) is reflected in scoring adopting the item response theory.

## A WORD TO CONCLUDE

The paper has addressed all the topics under interest. Substantially, TOEFL is a proficiency test which aims at assessing one’s general language ability. The test spreads individuals along the continuum of language ability so that their language abilities are known in an ability scale. The content of TOEFL does not reflect a particular set of syllabus or curriculum. Historically, TOEFL has witnessed three shifts in its version, namely pBT, cBT, and iBT along the line with a shift in the theory underlying their construction. The pBT and cBT are gradually being replaced by the iBT version, which began to come to public in 2005. Seen from the components making up TOEFL as a battery, each version has a different sub test with different testing formats. The earlier versions are characterized by the structural grammar views. The iBT version begins to move onto the communication-movement. The earlier ones make use of general social themes; the iBT version focuses more on academic matters. While the multiple-choice type still characterizes all versions dominantly, speaking and writing take 'direct' testing in which the test takers respond to the tasks by speaking and writing respectively not just recognizing alternative items provided as a selection. The scoring in all the versions changes moving from a complex score transformation to a simpler one.

As a proficiency test aimed at testing language abilities as are required in the academic settings, TOEFL has significant backwash impacts. For example, it has directed individuals to make a variety of attempts on how to achieve a higher TOEFL score. TOEFL training courses have mushroomed as a sequence. While all this is a good indicator of the presence of motivation in the English learning, cautions should be exercised. Tests like TOEFL frequently play a high-stake role. However, real academic life after taking TOEFL is more realistic. Therefore, a high score in TOEFL needs to reflect a solid mastery for functional communication in academic settings. The challenge for the TOEFL course providers is that they need to provide their customers with relevant instructional materials, suitable class tasks for practice, and more importantly they need to keep up with all recent advances made by ETS as the TOEFL developer.

## REFERENCES

Baker, Frank B. 1985. *The Basics of Item Response Theory*. Portsmouth, New Hampshire: Heinemann.

Brown, James D. 2005. *Testing in Language Programs: A Comprehensive Guide to Testing language Assessment*. New York: McGraw-Hill.

ETS, Undated. *TOEFL Internet-Based Test: Score Comparison Tables*. Princeton, New Jersey: Educational Testing Service.

Harris, David. P. 1969. *Testing English as a Second Language*. New York: McGraw-Hill Book Company.

Hulin, Charles L., Drasgow, Pritsz, and Parsons, Charles K. 1983. *Item Response Theory: Applications to Psychological Measurement*. Homewood, Illinois: Dow Jones-Irwin.

Jenskins-Murphy, Andrew. 1981. *How to Prepare TOEFL*. New York: Harcourt Brace Jovanovich.

Lord, Frederic M. 1980. Applications of Item Response Theory. Hillsdale, NJ: Lawrence Erlbaum Associate, Publishers.

Sharpe, Pamela J. 2005. *Barron's Practice Exercises for the TOEFL 5th Edition*. Jakarta: Bina Rupa Aksara.

Sulistyo, Gunadi H. (2001). 'Technical Considerations for Taking the (Paper-and-Pencil-Based) TOEFL'. A Paper Presented in a seminar *Computer-Based TOEFL: Concepts and Strategies* organized by CSU, English Department, State University of Malang, May 26, 2001.