

THE TOEFL EXAM: HOW RELIABLE IS IT?

By Denis Griffin
ITESM, Campus Zacatecas

INTRODUCTION

Due to the fact that Mexico is the only developing country in the world (with the exception of ex-communist countries of Eastern Europe) that shares a border with a developed country, and in this case the world's most powerful country, the United States has always exerted enormous influence on the Mexican education system in both the public and the private sector. One avenue of social advancement open to Mexican students can be initiated by gaining a scholarship to study in the United States or Canada. Another way of increasing employment status is to demonstrate to prospective employers that the candidate has a high or acceptable level of English. If students are to be seriously considered in either of these options of social and economic advancement, they usually must demonstrate their knowledge of English by taking a proficiency test. In the United States the official measure of a student's proficiency in English is the Test Of English as a Foreign Language (TOEFL). Many Mexican tertiary institutions also require it as a means of academic placement within their institution or as a requisite for graduation. A low score in the TOEFL will mean that the student will have to begin in a remedial level of English and need to work his/her way up slowly to an advanced level. Taking into account the powerful influence of this exam, one would expect that students could depend on a method of evaluation that will reliably and validly measure their level of English. Surely an exam that has involved so many language specialists in its formation and continual application, which has the official approval of the world's most powerful government and which is an export that generates millions of dollars annually for the United States, could be depended on for accuracy. The research that has been carried out for this paper will provide the reader with sufficient evidence to doubt the precision of the TOEFL as a means of measuring

English proficiency.

This study involves a comparison of students' results in the Institutional Testing Program TOEFL (ITP TOEFL) with the performance of the same students in an exam elaborated by the author. The exam written by the author requires open-ended answers. The author's hypothesis is that the format of multiple-choice options in the TOEFL comprises an artificial technique for examining students' ability in English and this indirect form of testing students does not represent a reliable means of evaluating their overall command of English.

LIMITATIONS OF THE TOEFL EXAM

Firstly, the limitations that the TOEFL exam presents in comparison to an intrinsically motivating exam based on the direct testing of language skills and sub-skills will be reviewed. The TOEFL is an exam composed of four parts: a listening section with 50 multiple choice questions that takes some 30 minutes or more; a structure and written expression section with 40 multiple choice questions that takes 25 minutes; a reading section with 50 multiple choice questions that take 55 minutes; and a written composition section that takes 30 minutes. The ITP TOEFL is an exam that does not include the last written section and is provided to institutions and then sent to TOEFL centers for evaluation.

The fact that the TOEFL exam now includes a written section is a big improvement over the previous format, which relied purely on multiple-choice questions, even if the manner of interpreting results in this section is questionable. Scorers of the TOEFL written exam have just one and a half minutes for evaluating each composition (Hughes, 1989, p. 86). The ITP TOEFL, however, has no such direct component for scoring a students' proficiency in writing. The ITP TOEFL is not officially recognized as the equivalent of the TOEFL, however, because many Mexican institutions use it for student placement and as a requirement for graduation, tertiary institutions in the United States and Canada often accept the scores of the ITP TOEFL, thus recognizing grades awarded in

Mexican institutions. Therefore, it could be said that the ITP TOEFL exerts an important influence within Mexican institutions and often decides whether students will be accepted into foreign universities.

The very fact that the ITP TOEFL is a two-hour exam based on only one kind of question (multiple-choice) makes it a long and tedious exam. Students must also race against the clock combining monotony with a certain level of stress. The last section is particularly grueling, as students have to tackle some five academic reading sections in a 55-minute period. This appears to be a brief period to accomplish these readings and answer 50 questions even for a native speaker. Obviously, not all students are mentally equipped to think quickly and choose from 4 different options constantly over a two-hour period. Furthermore, apart from the extrinsic motivation of obtaining a good evaluation, some students will feel very little intrinsic motivation to do well in such a monotonous exam, while others appear to just give up along the way. Such a situation in which a person is tested in this manner is rarely encountered in real life and those students who have ability in these kinds of exams would correspond to a particular intellectual profile.

H. Douglas Brown (1994) has proposed four principles to create intrinsically motivating tests. These are the principles of: giving students advance preparation; of face validity; of authenticity; and of "washback" (beneficial backwash) (pp. 385-387). Among the elements that give a test face validity, Brown mentions: "tasks that are familiar, that relate to their coursework"; "a difficulty level that is appropriate for (your) students"; and "test conditions that are **biased for best** - that bring out students' best performance" (Brown, 1994, p. 385). These three elements of face validity appear to be absent in the TOEFL as are also the last two principles mentioned above.

Such exams can also be said to be culturally biased and adapted to a very traditional Anglo Saxon concept of intelligence; a concept which has been criticized and disproved by educators and psychologists in recent years.

Traditionally in the West, intelligence was viewed as the ability to perform linguistic and logical-mathematical problem solving. Relating our traditional view of intelligence to the formulation of tests, Brown has argued that "since "smartness" in general is measured by timed, discrete point tests consisting of hundreds of little items, then why shouldn't every field of study be so measured? So, today we live in a world of standardized, norm-referenced tests that are:

timed
multiple choice
tricky
long
artificial" (1994, pp. 376-377).

This comment by Brown could be an adequate description of the TOEFL. Howard Gardner, however, has extended the traditional view of intelligence to seven components. Two of these components which are related to language learning but are ignored by traditional views are interpersonal intelligence and intrapersonal intelligence (Brown, 1994, p. 377). Robert Sternberg has also recognized peoples' creative thinking and manipulative strategies as part of intelligence. Brown expounds on this by stating that

"all "smart" people aren't necessarily adept at fast, reactive thinking. They may be very innovative in being able to think beyond the normal limits imposed by existing tests, and may need a good deal of processing time to enact this creativity. And other forms of smartness are found in those who know how to manipulate their environment. Debaters, politicians, successful salespersons, "smooth" talkers, and con artists are all smart in their own manipulative way" (Brown, 1994, p. 377)

Taking into account that tests like the TOEFL are geared to our traditional concept of intelligence and that students in Western institutions have been trained to tackle these kinds of exams, it is no wonder that Mexican students who come from smaller and more traditional cities and towns may have difficulty in scoring well on the TOEFL, even though they may make adequate progress in

their English courses.

In his insightful work "Testing for Language Teachers", Arthur Hughes (1989) makes a list of different testing criteria, the adopting of which can help teachers to achieve beneficial backwash, the beneficial effect of testing on teaching and learning (p. 1). Hughes mentions: testing the abilities whose development you want to encourage; sampling widely and unpredictably; using direct testing; making testing criterion-referenced; basing achievement tests on objectives; ensuring the test is known and understood by students and teachers; and providing assistance to teachers where necessary (pp. 44-47). Those elements most relevant to this particular study will be commented upon.

Firstly, in regards to testing the abilities whose development you want to encourage, Hughes states that "there is a tendency to test what it is easiest to test rather than what it is most important to test" (p. 44). This is especially true in relation to the TOEFL. The TOEFL is a particularly convenient exam to administer as the multiple choice answers can be easily scanned by a computer and a neat exact score is quickly calculated. In short it is a cheap and efficient way of generating income for the U.S. economy and of measuring supposed English ability. Hughes further argues that "when testing is carried out on a very large scale, when the scoring of tens of thousands of compositions might seem not to be a practical proposition, it is understandable that potentially greater accuracy is sacrificed for reasons of economy and convenience. But it does not give testing a good name! And it does set a bad example" (1989, pp. 2-3). Hughes continues by pointing out that "when we compare the cost of the test with the waste of effort and time on the part of teachers and students in activities quite inappropriate to their true learning goals, we are likely to decide that we cannot afford not to introduce a test with a powerful beneficial backwash effect" (1989, p. 47). Whereas reading, writing and listening are tested in the TOEFL, there is no oral test section included. This is one of the four major skills and therefore the TOEFL can be said to test indirectly or directly only 75% of the major skills involved

in English. One vital skill, which is included in other more authentic exams such as the Cambridge First Certificate, a British exam that incorporates more skills and testing techniques, has been completely omitted. Oral ability is also the language skill which is most noticeable, arguably the most practical and which is most judged by native speakers in determining the mastery that a second language learner has achieved.

Hughes' second element in achieving beneficial backwash is sampling widely and unpredictably. It is important that the sample being tested should represent as completely as possible everything included in the test specifications. The TOEFL writing test only sets two kinds of tasks: compare/contrast and agree/disagree (Mahnke and Duffy, 1996, p. 322). Preparation for this section of this test will, therefore, be limited to preparing students for completing only these two kinds of tasks, therefore beneficial backwash, although greatly enhanced from when there was no direct testing of writing, will still have a limited effect.

Direct testing also has an important beneficial backwash and is an essential element. Texts and tasks in testing the performance of skills should be as authentic as possible. The structure and written expression section of the TOEFL is an indirect method of testing writing ability. Hughes rightly argues that "immediately we begin to test indirectly, we are removing an incentive for students to practice in the way we want them to" (1989, p. 45).

Although the TOEFL is a criterion-referenced exam in that it provides students with a final score that has a maximum of about 677, there is no set passing or failing score. The examinee handbook for the ITP TOEFL states that "each institution determines for itself what scores, or ranges of scores, are acceptable.... There is no specific passing or failing score for the ITP TOEFL or Pre-TOEFL test" (ETS, 2001, p. 21). Hughes argues for a

"series of criterion-referenced tests, each representing a different level of achievement or proficiency. The tests are constructed such that a 'pass' is obtained only by completing the great majority of the test tasks successfully. Students take only the test (or tests) on which they are expected to be successful. As a result, they are spared the dispiriting, demotivating experience of taking a test on which they can, for example, respond correctly to fewer than half of the items (and yet be given a pass)." (1989, pp. 45-46).

This is one negative aspect of the TOEFL that must prove to be quite demotivating for English students, particularly when they know that they can only correctly choose less than 50 percent of answers, but still have the possibility of getting 100 percent in their course or at least a pass mark.

A final and major limitation of the TOEFL exam is the dominating use of multiple-choice questions. Multiple-choice exams, however, do have some advantages. They make scoring reliable due to their limited options, also rapid and economical. Due to the fact that the candidate has options and only has to make a mark on the examination answer sheet it is possible to include more items in the test and when a test has more items it can be more reliable (pp. 36-37, 59-60).

Hughes has analyzed different techniques for testing languages and provides a list of 6 difficulties associated with multiple-choice questions (pp. 60-62). Firstly, multiple-choice questions may only test recognition knowledge, whereas this kind of test may be a poor indicator of a students' ability to use, for example, grammatical structures. A person may be able to identify a correct response, but be unable to produce this structure in speaking or writing. Secondly, guessing may have a considerable but unknowable effect on test scores. In the TOEFL there is a 25 percent chance that they will identify the correct answer, which doesn't exist in open-ended answers. Multiple-choice items may also severely restrict what can be tested and it may be difficult to write successful items for these kinds of tests. If an institution depends on multiple-choice based tests to

evaluate the level of English in its campus in comparison to other campuses, backwash may be harmful. Often to practice for multiple choice items, as much time is spent on the practice of educated guessing as on content, and this may hinder a students' growth in language learning. Finally, one serious difficulty associated with multiple-choice exams is that they facilitate cheating. It is very easy to cheat off another student in the TOEFL simply by looking at the pattern of the circles that they have filled in. By facilitating cheating these kinds of exams may also hinder a students' development by placing them in a level that they are incapable of passing. The author is familiar with many students who were placed in higher levels by the ITP TOEFL, but who clearly did not have the language skills necessary to pass. Some admitted to have cheated in their ITP TOEFL placement test or in their preparatory ITP TOEFL exam previously.

A MORE AUTHENTIC ENGLISH PROFICIENCY EXAM

In order to test the reliability of the TOEFL to evaluate a student's proficiency level, firstly an exam was designed and administered to 39 students in two classes of "Lengua Extranjera", which is the advanced level of English at university level in the ITESM, Campus Zacatecas. The exam was designed to resemble the TOEFL, except that it had open-ended charts to be filled out for the listening section, cloze type questions (modified cloze and C-Test) for the structure and written expression section and open-ended questions for the reading section. The listening section was composed of two listening exercises consisting of 20 test items in total. The structure and written expression section consisted of 40 items, the same number as the TOEFL exam and the reading section of 10 questions based on only one reading. The reading section was probably the most unreliable section of this exam as it was composed of only one fifth of the number of questions to be found on the TOEFL. The exam was scored in the same way as the TOEFL also, because the correct sections on the listening section were multiplied by 2 and a half to equal the number of questions found on the TOEFL

(50), while the reading section was multiplied by 5 to represent the total of 50 questions. The structure and written expression section had the same number of questions (40). The number of correct answers was then converted into a score using a TOEFL conversion table so that a score like the TOEFL could be calculated. The results were then compared to the ITP TOEFL results.

As the research project progressed, however, it was found that the results from the first exam developed by the author were not very conclusive due to at least two reasons. Firstly, the reliability of this exam could be questioned due it having less test items than the ITP TOEFL and, secondly, the ITP TOEFL tested indirectly or directly more language skills (e.g. idioms, vocabulary, synonyms, conditionals etc.) in its three sections. Therefore the author decided to include two other ways of testing the reliability of the ITP TOEFL. To begin with two ITP TOEFL exams were administered to students within a two week period and their scores were compared. Later the author converted an exam with an identical format to the ITP TOEFL, which is used for practicing for the ITP TOEFL, into an exam with open ended answers. Later the same exam will be given section by section to the same students but in the typical multiple-choice TOEFL format and its results will be compared with the open ended exam. This part of the research project is still in progress.

THE RESEARCH PROJECT RESULTS

This project was comprised of different stages and cannot be said to be entirely finished. In the first stage the results from the ITP TOEFL and the exam designed by the author were calculated using a conversion table found in a TOEFL text used to teach the different skills required for this exam (Mahnke and Duffy, 1996, p. 505). At first glance the students in these two advanced groups performed poorly on the ITP TOEFL. The students had actually performed much better in the exam designed by the author.

The fact that the averages between the two exams differed so much made it

difficult to compare results, but at least it was interesting to compare the position that students finished in comparison to others. The two leading students in the ITP TOEFL actually had a similar score in both exams and their skills in different areas were similar, but many of the students received very different results. One of the biggest differences was the case of one student who finished in 32nd place in the ITP TOEFL, but in 2nd place in the second exam. According to her language ability observed by the teacher during the semester, the second exam seemed more reliable, although not conclusive. Some students seemed to have more talent for taking the type of multiple-choice exam represented by the ITP TOEFL, although their overall knowledge of English did not seem to be superior to the better students in the two classes.

A statistical analysis was made between the two exams to analyze their correlation. Firstly the correlation coefficient between the different sections of the two exams revealed a good correlation between the listening sections, a fair correlation in structure and written expression and a bad correlation in reading. The correlation coefficient between the percentage of correct answers in the different sections of students in the TOEFL and their points score in the TOEFL ranged between very good in listening and reading to good in structure and written expression. In the author's exam, the correlation coefficient in the listening and structure and written expression section was very good, while it was just fair in reading. As stated earlier, the reading section should have had more items and the lack of items probably affected the reliability of results. The standard deviation in the TOEFL was 43.84 points, which was less than the author's exam, which had a standard deviation of 48.62. The TOEFL also proved to have less dispersion than the second exam.

Next the author decided to examine the history of TOEFL scores for selected students who had taken the exam at least twice with the author. This data covers results from four semesters: from November 2000 to May 2002. Normally to stay within the passing score for a level, a student is expected to improve his score by

around 30 points from one semester to the next. Strangely enough, the scores of some students had actually fallen. From the teacher's points of view it was almost impossible that these students knew less English than they did in previous semesters. Only one student in the list actually increased his TOEFL score by an average of at least 30 points a semester, whilst the majority of students had increased their scores by a very modest amount.

Referring to reliability, Hughes has said that "a test is reliable if it measures consistently. On a reliable test you can be confident that someone will get more or less the same score, whether they happen to take it on one particular day or on the next; whereas on an unreliable test the score is quite likely to be considerably different, depending on the day on which it is taken" (1989, p. 3). Taking this into account two ITP TOEFL exams were administered to students in within a two week period and the results compared. Whereas the class average changed very little, individual scores changed dramatically with some rising or falling more than 50 points, the equivalent of almost two levels in the ITESM system.

The author is still working on the latest stage in the research project, that of comparing the results of an exam with an identical format to the ITP TOEFL with the results of the same exam with open-ended answers.

CONCLUSIONS

Recent research and scholarship have questioned the reliability and validity of exams based on multiple choice questions such as the TOEFL and the ITP TOEFL. While there was some correlation between the ITP TOEFL results and the first exam designed by the author, the correlation for the different test sections ranged between good and bad. The teacher of these two groups, however, suspected that the author designed exam may have contained more reliability pertaining to the overall English level of the groups. Moreover the history of TOEFL scores of students who have taken this exam more than once with the author in the two "Lengua Extranjera" groups suggest that ITP TOEFL results can be quite misleading from one semester to another. A later stage of the research project revealed that scores from two ITP TOEFL exams for many individual students rose and fell considerably within a two week period. The latest stage of the project is still in progress. Overall, the results so far in this project indicate that ITP results don't necessarily provide academic institutions with an accurate account of students' overall mastery of English.

REFERENCES

- Broukal, Milada (1995). *The Heinle & Heinle TOEFL test assistant*. Boston: Heinle & Heinle.
- Brown, H. Douglas (1994). *Teaching by principles: An interactive approach to language pedagogy*. Englewood Cliffs, NJ: Prentice Hall Regents.
- Educational Testing Service (ETS) (2001). *TOEFL institutional testing program: Examinee handbook and admission form*. Princeton, NJ: ELS.
- Harmer, Jeremy (1991). *The practice of English language teaching*. London: Longman.
- Hughes, Arthur (1989). *Testing for language teachers*. Cambridge: Cambridge University.
- Mahnke, M. Kathleen, & Duffy, Carolyn B. (1996). *The Heinemann ELT TOEFL preparation course*. Oxford: Macmillan Heinemann.
- Maurer, Jay (2001). *Focus on grammar: An advanced course for reference and practice. Teacher's manual*. White Plains, NY: Longman.
- Phillips, Deborah (2001). *Longman complete course for the TOEFL test: Preparation for the computer and paper tests*. White Plains, NY: Longman.
- Purpura, James E., & Pinkley, Diane (2000). *On Target 2 teacher's edition: Intermediate*. White Plains, NY: Longman.