PASAA
Volume 41
November 2007

Are They Making an Appropriate Decision? Implications from Bachman's Assessment Use Argument

Yi-Ching Pan
National Pingtung Institute ofCommerce, Taiwan
Email: huangpan63@yahoo.com

Abstract

Tests do not occur in a vacuum; they have far-reaching and unanticipated consequences. Tests have an effect not only on traditional teaching practices in the classroom but also on the broader educational and societal contexts. Examples are the selection of candidates for education, the monitoring of the performance of schools, the distribution of funding and even societal issues such as the selection of employees and candidates for immigration, citizenship and asylum. Given the potential power of tests, it is essential to justify test use and investigate its consequences. The primary goal of this paper is to utilize Bachman's assessment use argument as a model for evaluating the appropriateness of test use. This paper begins with the issue of the importance of the justification of test use when validating a test. Next, Bachman's model is illustrated. Following that, Bachman's model is adopted to investigate whether it is appropriate to use the GEPT (General English Proficiency Test) as a graduation threshold in Taiwan. In order to do this, both the warrants for and rebuttals against using the assessment for this decision are discussed. This paper provides logical arguments for test users and also suggests what evidence needs to be collected to make an informed decision regarding establishing an English requirement as a graduation threshold. It is hoped that this paper will serve as an example of what needs to be taken into account when making the decision of test use in order to make the decisions convincing and therefore beneficial to those affected.

Key Words: consequences of test use, an English requirement for graduation, validity, washback

Introduction

In Taiwan, English is taught as a foreign language (EFL) within a classroom-based environment. Officially, students begin learning English in the fifth grade, but in large urban areas such as Taipei, Taichung, and Kaohsiung, children usually learn English earlier, normally from the first grade on. After at least two years of English instruction in elementary school, students will receive six years of English education before they attend colleges or universities. Students need to take two public exams, one Basic Competence Test (BCT) to enter senior high school and the College Entrance Examination (CEE) for higher institutes of learning. These two examinations evaluate students' English proficiency, and their English scores are taken as one of the criteria for school admissions and used by students to help choose the school they wish to attend. University students are usually required to take 3-4 hours of English every week in their first year.

Despite significant exposure to English (nine years of English classes from elementary school to college/university), the TOEFL (Test of English as a Foreign Language) CBT Score Data Summary from 2002-2006 provided by the Educational Testing Service (http://www.ets.org/Media/Research), shows Taiwanese students' scores ranked from the fourth-lowest to the seventh-lowest among the thirty-two countries in Asia. In another ETS survey done in conjunction with National Chengchi University in Taiwan, 32.3% of Taiwan's college students examined for English proficiency function at the level of students in their third year of junior high or first year of high school (Huang, 2003). According to its developer, the LTTC

(Language Testing and Training Centre), the GEPT (General English Proficiency Test, for more detail, see the following section) is considered competency equivalent to a junior high student's English proficiency. However, the percentage of college graduates who have passed the first stage of the GEPT elementary level, based on the LTTC score statistics in 2002 (http://www.ltc.com.tw), was only 14.9%.

In order for their students to attain a certain level of English proficiency and equip them with higher competitive strength in the job market, an increasing number of universities and colleges in Taiwan have set an English proficiency requirement as a determining factor for an individual's readiness to graduate.

Another reason for setting the exit requirement is to facilitate the government's educational policy. Taiwan's government is aware of the importance of English and considers the lack of English proficiency in it problematic because the English language is now regarded as an essential worldwide communication tool. Therefore, the government seeks to develop national proficiency in order to be more competitive in global markets. As part of endeavour, for example, in 1999, the Ministry of Education commissioned the Language Testing and Training Centre to develop the General English Proficiency Test (GEPT) to "promote lifelong learning and encourage the study of English" (http://www.lttc.com.tw). Since 2003, the Ministry of Education has encouraged universities and colleges of technology to set English thresholds for graduates so that they will be able to achieve a level of proficiency to meet the anticipated needs of both domestic and international job markets. Moreover, a priority goal of the four major educational policy pivot points for 2005-2008 proclaimed by the Ministry of Education in February 2004 is to have 50% of students at universities and colleges of technology achieve an English proficiency equivalent to General English Proficiency Test (GEPT) Intermediate and Elementary Levels, respectively, by 2008. The government has also provided funding to assist colleges to reach that goal.

Some universities and colleges of technology have adopted the GEPT or other English proficiency tests such as TOEIC, TOEFL, CSEPT (College Student English Proficiency Test) and school-designed tests as graduation thresholds. Some require students to pass GEPT Intermediate or Elementary Level, TOEFL CBT at 193 (or TOEFL paper and pencil test at 500), CSEPT (College Student English Proficiency Test) Level 1 or 2 or school-designed proficiency tests. Other universities and colleges that have not established English exit requirements have set reward policies to encourage students to pass English proficiency tests by offering them financial incentives or waiving compulsory English classes.

Graduation thresholds requiring English proficiency tests, however, have met with opposition. Schools such as Ming Chuan University in Taipei (Zhang, 2005) argue that universities and colleges are not cram schools. They do not promote "teaching to the test", so they do not set English graduation thresholds. Instead, they require students to take more English-related classes to enhance their proficiency. Some English educators hold similar opinions. Liao (2004) is concerned that English proficiency requirements will force teachers to teach to the test because curricula will mirror exam content. Negative washback from the requirement will most likely manifest itself in teachers teaching to the test, students cramming for tests, and the narrowing of curricula. Liao suggests that students should be immersed in an English environment. Setting English proficiency test requirements for graduation is not a panacea. Some legislators (Mang, Zhang, & Lin, 2003) have also expressed their objections. Legislator Li Chin Ann contends that there is little point in setting such thresholds because students are already required to take regular English classes. As long as students pass these classes, why should they have to take English proficiency tests?

The purpose of this paper is to explore whether it is appropriate to establish English proficiency as a requirement for graduation, focusing on using the GEPT as the primary standard for graduation given its overwhelming predominance in the field of proficiency tests among students in Taiwan, as compared to others such as TOEFL, TOEIC, and IELTS. To address this issue, I will begin with an overview of the General English Proficiency Test (GEPT). Next, I will discuss the importance of evaluating test use when validating the test. Following that, I will present an introduction to Bachman's Assessment Use Arguments (2005) will be presented. Following this, one part of Assessment Use

Arguments, Assessment Utilization Arguments, will be adopted to investigate the GEPT as a graduation threshold. In order to do this, both the warrants for and rebuttals against using the assessment for this decision will be discussed. This paper will provide logical arguments for test users and also suggest what evidence will need to be collected to make an informed decision regarding establishing an English requirement as a graduation threshold. It is the researcher's hope that this paper will serve as an example of what needs to be taken into account when making the decision of test use.

An Overview of the General English Proficiency Test (GEPT)

In 1999, to promote the concept of lifelong learning, to further encourage the study of English, and to offer students of English a fair and reliable test (http://www.lttc.ntu.edu.tw), the Ministry of Education in Taiwan commissioned the Language Training and Testing Center (LTTC) to develop the General English Proficiency Test (GEPT). Based on the information on the LTTC website (http://www.ltc.ntu.edu.tw), the GEPT is used by individuals to assess their proficiency levels, by institutions both private and public to serve as a placement and promotion criterion for employees and by some institutes of higher education as an admission or graduation requirement. Since the GEPT began being administered in 2000, it has been the subject of much public interest. To date, more than 1.74 million test-takers have taken it.

The GEPT tests at five levels: Elementary, Intermediate, High-Intermediate, Advanced, and Superior. It is designed to test four skills: listening, speaking, reading, and writing. Test-takers can choose the level of the GEPT best suited to their proficiency. They must pass two stages of the exam in order to receive GEPT certificates. Test-takers who pass the first stage have two chances to take stage 2 should they fail it the first time.

The first stage tests for listening and reading comprehension at all levels except the Elementary, which includes listening, reading and writing for the first stage. All the questions except for writing are multiple choice. For the Elementary, Intermediate and High Intermediate Levels, the speaking test is given in a semi-direct way. Test-takers, after hearing a tape, record their answers by reading aloud, answering questions, and giving picture descriptions. At the Advanced and Superior Levels, a direct speaking test is given in the form of interviews, information exchange, and presentations by interacting with an interlocutor. For the written tests, the questions vary from sentence-combining and sentence-making to short-essay writing. Details of test format, structure, and time allotment for each level is given in Appendix 1. The general skill description for each level is given in Appendix 2.

As an example, at the GEPT Elementary Level, each part includes rubrics spoken in Chinese to explain to test-takers what they will be tested on in terms of four skills: listening, reading, writing and speaking

1. Listening: This section has three parts, each of which is multiple-choice. The first part is picture description, which asks students to answer questions they hear on tape about an image they observe. The second is either one short question or statement to which they are expected to choose the best response or meaning from four answers. In the third part, they first listen to a short conversation then answer a question based on the conversation. For parts I and II, students only listen once, but for part III, they hear each short conversation twice. The questions reflect the language necessary in daily life for such topics as prices, times, places, foods, and transportation. Test-takers have about 20 minutes to answer 30 items.

2. Reading: This section has three parts, and all of them are multiple-choice. The first features questions about vocabulary and sentence structure. The second is a close test, in which test-takers are required to choose the most appropriate word or preposition to fit in the context. The third part is reading comprehension. Most of the questions in the reading section test students' vocabulary, grammar and reading abilities. The reading questions test knowledge regarding concepts such as street and traffic signs, shop signs, simple menus, schedules and greeting cards. Test-takers have about 35 minutes to answer 35 questions.

3. Writing: This section has two parts. The first part asks test-takers to write sentences by combining two , transform them by varying their tenses, voices, and purposes (declarative, interrogative, etc.) and so on. The second requires a written 80-100-word essay based on the picture given. The time allotment for this section is about 40 minutes.

4. Speaking: This section has three parts. The first involves repetition, that is, test-takers are required to repeat a sentence they hear on tape. Each sentence is read twice. In the second, test-takers read aloud. They are given a short paragraph, and after a minute to look over the transcript, they read the paragraph into the recorder. For the third part they listen to recorded questions twice then speak their answers back to the recorder. Most questions are related to real-life use such as greetings, school life, and asking for directions. The time limit for this section is about 10 minutes.

5. Scoring: In the first stage, the passing score for listening and reading is 80 out of 120 . The passing writing score for Elementary Level at this stage is 70 out of 100. In the second stage, the passing score for writing is 80 out of 100. Writing and speaking are scored by trained raters on a holistic scale from 0 to 5 then converted to percentile points.

The Importance of Evaluating Test Use When Validating a Test

Tests influence both education and society—as Bachman (1990) states, tests are not developed and deployed in a "value-free psychometric test-tube" (p. 279) and they are intended to fulfill the needs of an educational system or society. Shohamy et al. (1996) share a similar belief that testing does not occur in a vacuum in that the results elicited can have far-reaching consequences both on individuals and programs.

From a micro point of view, it has been shown, over and over, that tests affect the behavior of teachers and students as well as how they perceive their individual performance and value. Tests may also determine educational content and methodology (Wall, 1997). Pearson (1988) points out that testing exerts an influence over both teachers and students in attitude, behaviour and motivation.

From a macro point of view, tests can have consequences not only within the classroom but also on the educational system and society as a whole. Cheng (2005) and McNamara & Roever (2006) provided several examples: the use of examinations for selecting candidates for education, employment, promotion, immigration, citizenship or asylum, monitoring the performance of schools and colleges, implementing educational policies, reforming educational systems, deciding on the distribution of funding, etc.

Given the consequences that tests exert at both micro and macro levels, researchers (Bachman & Palmer, 1996; Bachman, 1990; McNamara & Roever, 2006; Messick, 1989; Messick, 1996) argue that when creating a test, focus solely on the investigation of content, criterion-referenced and construct validity is insufficient; researchers should also consider the consequences of test use. Therefore, in the last twenty years, the concept of test validity has become increasingly broad and complex, dependent on data derived from a range of evidence sources rather than solely from the validation of the test itself. Included in this wider evidence base are the purposes and circumstances of test use.

In his acclaimed paper regarding validity, Messick (1989) developed the concept of consequential validity—the influences of score interpretation and test use. The concept of washback is connected to the validity of the test and associated with Messick's consequential validity. Messick (1996) viewed washback as an "instance of the consequential aspect of construct validity" (p. 242), which covers elements of test use, the impact of testing on test-takers and educators, the interpretation of results by decision-makers, and any possible misuses, abuses, and unintentional uses of tests (Messick, 1989). Many other researchers (Bachman, 2005; Cronbach, 1988; McNamara, 2006; McNamara & Roever, 2006; Shohamy, 2001) have also stressed the importance of justifying test use and investigating its consequences.

In light of the foregoing, validity has shifted focus from the wholly technical viewpoint to that of a test-use perspective (Mousavi, 2002). As Kane (2001; 2002) has pointed out, test validation at this stage is not to examine if the test itself is valid or if test scores per se are validated, but rather that the interpretation, inferences or decisions of test use are subject to validation. Validity is taken as a unified point of view so that a collection of techniques is simply not enough, and validation is an ongoing process of providing a variety of evidence about test interpretation and use (Bachman, 1990). McNamara and Roever (2006) also state that since tests can have widespread and unforeseen consequences, a language test that is psychometrically validated does not necessarily denote a test favorable for society. The researchers then propose the need to develop a social theory to assist test developers and researchers in better comprehending testing as a social practice for their work. To conclude, other than evaluating psychometrically based analyses of score meaning, validity has also drawn attention to investigating test use, particularly on the issue of its consequences.

Bachman's Assessment Use Argument

Before the 1980s, "validity was described as a characteristic of a test: the extent to which a test measures what it is supposed to measure" (Chapelle, 1999, p. 258). To investigate the validity of tests, content validity, criterion-related and construct validity are usually utilized. In addition, establishing validity was regarded as the responsibility of testing researchers when developing large-scale, high-stakes tests. Validity is currently viewed as an argument concerning test interpretation and use: the extent to which test interpretations and uses can be justified. Validity is a unitary concept with construct validity at its core. Content and criterion-related evidence can be used as evidence for construct validity. Meanwhile, justifying the validity of test use has become the responsibility of all test users (Chapelle, 1999). From the aforementioned discussion about the different concepts of validity before and after the 1980s, validity shifts from the technical viewpoint to that of a test-use viewpoint (Mousavi, 2002). It then becomes an essential issue of what can be used to justify test use. Bachman (2005) contends that there is little literature in the fields of language testing that present a set of guidelines or procedures for connecting test scores and score-based inferences to test use and its consequences. Bachman also mentions that although Messick (1989) discusses test use and consequences, his framework doesn't depict how to examine the relationship between test scores and test use and the consequences of test use. In other words, Messick doesn't develop a comprehensive set of procedures to link score interpretations to test use and the consequences of test use. Likely, argument-based formulations of validity (Kane, 2001, 2002), based on Bachman's point of view, do not address the issues of test use and consequences of test use, although they do offer a set of procedures for investigating and supporting data regarding score-based inferences. In viewing the lack of a link between validity and test use and consequences of test use, Bachman has adopted Toulimin's model of "assessment validity argument" and also integrated Kane's interpretative argument, and then created his "assessment utilization argument". The two were combined and given a new name: Assessment Use Argument. Bachman hopes that the procedures in his framework will be beneficial for stakeholders, no matter whether they are test developers or test users, and will collect the most critical evidence in support of the interpretations, test uses and test decisions.

Bachman's "assessment use argument" consists of two parts, an assessment validity argument, which establishes a connection between assessment performance and its interpretation, and an assessment utilization argument, which establishes a connection between the interpretation and a decision.

An assessment validity argument is made up of four components:

Claim : interpretation as to the conclusion made based on a test-taker's test performance

Data : the information about how a test-taker performs on a test

Warrant : arguments made to support the claim

Backing : arguments made to support the warrant

Rebuttal : alternative explanations or counterclaims

Rebuttal data : arguments made to support, weaken or reject the rebuttal

In summary, an assessment validity provides a set of procedures on how to make a valid interpretation of scores. First, data gives information about what a test-taker knows or can do.

Next, warrant, which based on the supporting arguments, provides the criterion or rationale for the interpretation. Then, warrant should be supported by backing, drawn from theory, prior research or experience. Or there may be an alternative interpretation called rebuttal, which can be backed up, diminished, or abolished by rebuttal data.

According to Bachman, an assessment validity can be used to justify the score-based interpretation, and it can also be a necessary component for justifying test uses, but it is not sufficient. First, even valid score-based interpretations do not necessarily provide relevant, useful and sufficient information for test uses or decisions. Second, there may be some other test uses or decisions that will distort the original ones that interpretations have been intended for. Third, an assessment validity has no potential to predict or investigate unintended consequences of how the score-based interpretations are used. To conclude, the assessment validity argument is not able to link the interpretation and test use. In view of the limits of an assessment validity, Bachman proposed an assessment utilization argument to provide explicit links between score-based interpretations and decisions and test uses.

Like the assessment validity argument (see figure 1), the assessment utilization argument has the same structure, using claims, warrants, backing and rebuttals, but these terms will be a bit different from those in the validity argument. The terms will be explained in turn below:

Claim : the decision to be made

Data : score-based interpretations from the validity argument

Warrant : the information here has to be relevant, useful, beneficial and sufficient for making the decision

Backing : prior research, evidence, social practice and values, government regulations, laws and legal precedent, anything that can be used to support warrants

Rebuttal : alternative reasons to turn down an intended decision or to make a different decision; unintended consequences of using the assessment and/or making the decision

Rebuttal data : arguments used to support, weaken or even reject rebuttals

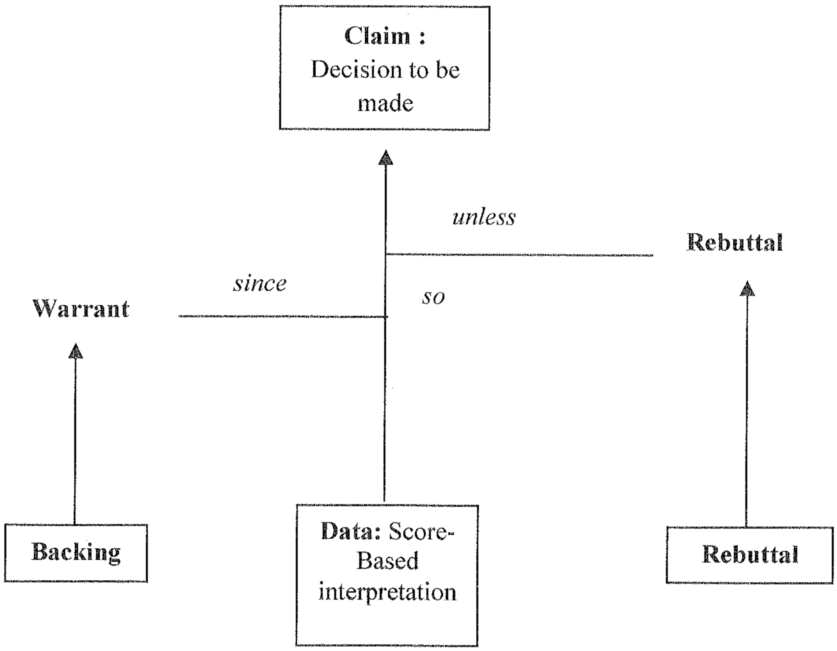

Figure 1 Bachman's Assessment Utilization Argument, from Bachman, 2005, p. 18

Assessment Utilization Argument as a Justification for the GEPT as a Graduation Threshold

The main purpose of this study is to investigate whether it is appropriate to set up an English requirement as a graduation threshold for students at colleges of technology, so assessment utilization argument will be utilized as the set of procedures to justify the decision. Both the warrants for and rebuttals against using the assessment for this decision will be discussed. By comparing the warrants and rebuttals, test-users thus will realize if they have made a right decision, or if they are going to make the decision, what they need to be aware of so that the decision is useful, valid, fair and ethical for test-takers.

1. Claim: Whether it is appropriate to set up an English requirement as a graduation threshold for students at colleges of technology.
2. Data: The GEPT scores
3. Warrants:

a. **Warrant 1 (relevance):** English is an important language for students no matter whether they will utilize it for further studies or for competitiveness in the job market, so it is crucial that students maintain a certain level of English proficiency.

Backing 1: English proficiency is the basic requirement for most business corporations when they recruit employees and for most graduate schools when they choose candidates. Both employees and graduate students need to master English at a certain level so that they then will be able to handle the tasks related to English, such as writing business letters or academic reports, answering international phone calls, making presentations either in an academic or business setting, reading up-to-date information regarding technology, and so on.

b. **Warrant 2 (utility):** Students' GEPT scores are excellent predictors of the requisite language ability necessary for English tests or interviews required of students by certain businesses and graduate schools.

Backing 2: To support this argument, test users need to do empirical evidence and find out if people with the GEPT certificates have received higher grades on English tests or interviews when they look for jobs or take graduate entrance exams. Or empirical evidence can be conducted to examine whether those who are admitted to graduate school or certain accredited business corporations do better on the GEPT than those who are not.

c. **Warrant 3 (intended consequences)**: By setting the English requirement as a graduation threshold, students will be encouraged to study harder, and it will not only enhance their English proficiency but also create an English-study ambition in the student. When most students are studying English, everyone else will also be motivated to study English. Moreover, the number of students who hold the GEPT certificates will help build up the school's academic reputation and ranking, and thus the school will be able to recruit more potential students who are interested in English.

Backing 3: To prove these arguments, research needs to be conducted to investigate if the students in the schools that set the GEPT requirement as a graduation threshold spend more time studying English, and if these schools are also ranked better than those that do not have such a graduation threshold.

d. **Rebuttal 4 (sufficiency)**: To date, the GEPT is acknowledged and accredited as an objective test to access students' English proficiency in four skills: reading, writing, listening, and speaking.

Backing 4: To support this assertion, questionnaires and interviews with/for students and teachers can be distributed to examine the face validity of the GEPT. Moreover, reports about the reliability and validity made by the LTTC, which has developed the GEPT, can be also used as evidence to convince test-takers, teachers, and even parents of the objectivity of using the GEPT requirement as the graduation threshold.

4. Rebuttals:

Rebuttal 1: Not every student can afford to take the GEPT.

Rebuttal data 1: Data needs to be collected to show the percentage of students who can't afford to sit for the GEPT. What policy should be enacted for students who cannot afford to take the GEPT but will also be fair to those who can afford to take the GEPT?

Rebuttal 2: The course curricula and syllabi may not be able to cover the four skills that are tested on the GEPT. The class size in Taiwan tends to be big, ranging from forty up to sixty students in one class. For training speaking and writing skills, one teacher is insufficient to take care of every student. Additionally, there may not be enough classes offered to enhance students' English proficiency. Normally, technical colleges require students to take three credit hours of English for two semesters; in other words, are a year of six-hour English classes sufficient to help students pass the certain level of the GEPT? Also, most of the required English classes in Taiwan focus more on reading skills, and this may not be enough to prepare students to acquire the four skills that are tested on the GEPT.

Rebuttal data 2: Investigation on course curricula and syllabi needs to be done in order to find out what teachers have been teaching in class, and whether they also focus on enhancing English in four skills or on certain skills such as reading and listening. If what teachers instruct in class is not what the GEPT is intended to test, it's not fair to ask that students have to pass the GEPT in order to graduate. It's like wanting your horse to run fast but not providing it with any food, or expecting soldiers to fight a war for you without giving them guns or bullets.

Rebuttal 3: Teachers may not be aware of the test content on the GEPT, so they may not be ready to help students in preparation for passing the exam.

Rebuttal data 3: Teacher questionnaires, interviews with teachers and classroom observations can be conducted to determine teachers' knowledge on how and what is tested on the GEPT.

Rebuttal 4: Unexpected pressure and anxiety may be imposed on both teachers and students. In this sense, teachers may "teach to tests" and students "study to tests", and finally it may end up that students do not study English at all once they have passed the GEPT requirement.

Rebuttal data 4: Student and teacher questionnaires, interviews and classroom observations can be set to see the degree and the intensity of pressure and anxiety that teachers and students experience that is brought about by the GEPT.

Rebuttal 5: Students or teachers may think there is no need to take the GEPT as proof of English proficiency. Academic transcripts and grades from English courses can be used as an indicator of English ability as well. Passing the GEPT doesn't necessarily mean that students will get good grades in English courses, and sometimes it is the case that students who pass the GEPT fail their English courses.

Rebuttal data 5: The relationship between the GEPT scores and the subject grades can be investigated through correlation analysis.

Summary and Conclusion

Bachman's "assessment use argument" consists of assessment validity argument and assessment utilization argument. These two arguments provide a set of procedures for test developers and test users to follow when they justify the score-based interpretations and test uses/decisions. In the beginning, this study invokes a question whether it is appropriate to set the GEPT requirement as the graduation threshold. The answer now is still yes and no. Warrant arguments to support and rebuttal arguments to reject the decision for the GEPT requirements are discussed. Whether the decision is practicable or not, the suggested methods to find proof or evidence for "backing" and "rebuttal data" can be conducted and the findings from both quantitative and qualitative studies can be used as support for whether to make or reject the decision. If the test users decide to set up the GEPT requirement, then "backing" evidence should be found, and this study has provided certain directions for test users when looking for evidence. Conversely, if the test users decide to reject the GEPT requirement, this study also has offered significant rebuttals, and again how to find support is directed under the heading of "rebuttal data." This study does not mean to suggest whether to make or reject the GEPT requirement, but instead hopes to serve as an example for test-users and provide them with some suggestions, directions, and methodologies for addressing the issue of making an English requirement the graduation threshold. It is suggested that test users come up with more warrants with its backing evidence and rebuttals with its rebuttal data so that it will help them make a more appropriate decision and thus can be more accountable to test takers who are affected by the assessment and decision.

Acknowledgements

The author is grateful to Dr. Carsten Roever, senior lecturer at the Department of Linguistics and Applied Linguistics at the University of Melbourne in Australia for his advice and input into this study.

The Author

Yi-Ching Pan, who is a lecturer at the National Pingtung Institute of Commerce in Taiwan, has been awarded the Melbourne International Research Scholarship Pan and is taking her PhD at the Department of Linguistics and Applied Linguistics at the University of Melbourne, Australia. Her field of research is language testing with emphasis on washback in particular.

References

Bachman, F., & Palmer, S. (1996). *Language testing in Practice.* Oxford: Oxford University Press.

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

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. . *Language Assessment Quarterly, 2*(1), 1-34.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics,, 19,* 254-272.

Cheng, L. (2005). *Changing Language Teaching Through Language Testing: A Washback Study.* Cambridge: Cambridge University Press.

Cronbach, L. J. (1988). Five perspectives on validation argument. . In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Huang, H. M. (2003, November 6). The survey on college students' English proficiency. *Central News Agency.*

Kane, M. T. (2001). Current concerns in validity. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

Liao, P. (2004). The effects of the GEPT on college English teaching (全民英檢對於大專英語教學的影響). from http://home.pchome.com.tw/howbiz/posenliao/page/page_5_011.htm.

Mang, X. J., Zhang, J. H., & Lin, X. F. (2003, December 19, 2003). Legislators: The General English Proficiency Test cannot be set as the graduation threshold (全民英檢 立委：不得列畢業門檻). *United News (聯合報).*

McNamara, T. (2006). Validity in Langauge Testing: The Challenge of Sam Messick's Legacy. *Language Assessment Quarterly, 3*(1), 31-51.

McNamara, T., & Roever, C. (2006). *Language testing: the social dimension.* Oxford: Blackwell.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3).

Pearson, I. (1988). Tests as levers of change (or "putting first things first"). In D. Chamberlain & R. Baumgartner (Eds.), *ESP in the classroom: Practice and evaluation ELT Documents #128.* (pp. 98-107). London: Modern English Publication in association with the British Council.

Shohamy, E. (2001). *The Power of Tests.* Harlow: Pearson Education Limited.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: washback effect over time. *Language Testing, 13*(3), 298-317.

Wall, D. (1997). Impact and Washback in language testing. In C. C. & D. Corson (Eds.), *Encyclopedia of language and education. Language Testing and Assessment. Vol . 7* (pp. 291-302).

Zhang, Y. F. (2005, December 6). The General English Proficiency Test as a graduation threshold: the high-intermediate level for National Taiwan University and National Chengchi University (英檢畢業門檻台、政大中高級). *United News (聯合報).*

Appendix 1

The format and structure of the GEPT

(Source: LTTC Website: http:www.lttc.ntu.edu.tw)

Level Format	Elementary	Intermediate	High-Intermediate	Advanced	Superior
Listening	1.Picture description 2. Question or statement response 3. Short conversation (30 items) (20 minutes)	1. Picture description 2. Question or statement response 3. Short conversation (45 items) (30 minutes)	1. Question or statement response 2. Short conversation 3. Short talk (45 items) (35 minutes)	1. Short conversation or talk 2. Long conversation 3. Long talk (45 minutes)	No listening test for this level.
Reading	1. Vocabulary & structure 2. Cloze 3. Reading comprehension (35 items) (35 minutes)	1. Vocabulary & structure 2. Cloze 3. Reading comprehension (40 items) (45 minutes)	1. Vocabulary & structure 2. Cloze 3. Reading comprehension (50 items) (50 minutes)	1. Careful reading 2. Skimming & scanning (70 minutes)	No reading test for this level.

Level Format	Elementary	Intermediate	High-Intermediate	Advanced	Superior
Writing	1. Sentence writing 2. Paragraph writing (16 items) (40 minutes)	1. Translation 2. Guided writing (2 items) (40 minutes)	1. Translation 2. Guided writing (2 items) (50 minutes)	1.Summarizing & expressing opinions 2.Summarizing and providing solutions (105 minutes)	1. Activity 1: Listening 2. Activity 2:Reading 3. Writing task (3 hours)
Speaking	1. Repeating 2. Reading aloud 3. Answering questions (18 items) (10 minutes)	1. Reading aloud 2. Answering questions 3. Picture description (13-14 items) (15 minutes)	1. Answering questions 2. Picture description 3. Discussion (10 items) (20 minutes)	1. Warm-up interview 2. Information exchange 3. Presentation (25 minutes)	1.Presenta - tion 2.Answering questions (50 minutes)

Appendix 2

General Level Descriptions of the GEPT

(Source: LTTC website: www.ltt.ntu.edu.tw)

Level	Skill	Equivalent	Recommended Jobs
Elementary	a. Understand and use rudimentary language needed in daily life.	b. Roughly equivalent to the level of a junior high school graduate in Taiwan.	c. administrative assistants; maintenance personnel; taxi drivers; service personnel in department stores, restaurants, hotels and tourist facilities
Intermediate	a. Be able to use basic English to communicate about topics in daily life.	b. Roughly equivalent to the level of a high school graduate in Taiwan.	c. administrative, marketing, and sales personnel; technicians; nurses; hotel reception personnel; switchboard operators; police officers; tourism industry workers
High-Intermediate	a. Be able to have a generally effective command of English, and to handle and communicate a broader range of topics, although there might be some minor mistakes not influencing communication.	b. Roughly equivalent to the level of a university graduate in Taiwan whose major was not English.	c. business professionals; secretaries; engineers; research assistants; airline flight attendants; airline pilots; air traffic controllers; customs officials; tour guides; foreign affairs police; news media personnel; information management personnel
Advanced	a. Be able to communicate fluently with only occasional errors related to language accuracy and appropriateness, and to handle academic or professional requirements and situations.	b. Equivalent to the level of a university graduate who majored in English.	c. high-level business professionals; negotiators in business and government; English language teachers; researchers; translators; foreign affairs officials international news personnel

Level	Skill	Equivalent	Recommended Jobs
Superior	a. Be able to communicate effectively in all kinds of situations.	b. Equivalent to the level of a native English speaker who has received higher education.	c. Ability in this level is recommended for interpreters; overseas correspondents working for new agencies, foreign diplomats, high-level negotiators in business and government.