STANDARD SETTING FOR THE LISTENING SUBTEST ON A UNIVERSITY ESL PLACEMENT TEST

WEI-WEI YANG

University of Hawai‘i

INTRODUCTION

What is Standard Setting, and Why is it Important?

Standard setting is defined here as “the task of deriving levels of performance on educational or professional assessments, by which decisions or classifications of persons (and corresponding inferences) will be made” (Cizek, 2001, p. 3). The product of the standard setting task is a cut point or several cut points that correspond to the performance standards and separate people into different categories based on the assessment results. A *cut point* is also known as cutscore, cutting score, cut-off score, passing score, etc; and people can be categorized into pass/fail, certified/non-certified, beginning/intermediate/advanced, level 1/2/3, etc., depending on the number and the nature of performance levels involved.

Standard setting and the cut point(s) produced from the activity clearly affect peoples’ lives to different degrees and in various life situations. In most assessment situations, especially in high-stakes ones, an inappropriate or unfair cut point derived from the standard setting process may pose threats and cause harm to people’s financial condition, social status, health, psychological and mental state, etc. For example, an unreasonably high cut point for a high school graduation test may deprive many good students’ of the right to get their high school diploma. On the other hand, an overly low cut point in this case may grant many unqualified students a school diploma that they do not deserve. As another example, consider an unjustified and unreasonably low cut point for a doctor licensure test, which may certify candidates who are not ready to practice yet and are likely to misdiagnose or mistreat the patients, or an unreasonably high cut point that may rule out those who are ready and qualified. Neither policy makers

Second Language Studies, 26(1), Fall 2007, pp. 59-84.

nor ordinary people want the above scenarios to happen. Therefore, standard setting deserves proper attention and all necessary effort, through which we hope to obtain reasonable and defensible cut points that can truly reflect the performance levels and can put people into the “right” categories.

Development of Standard Setting in Educational Measurement

Measurement professionals started to pay attention to standard setting from the 1950s onward, and a number of influential methods for standard setting were invented between the 1950s and 1970s (e.g., Nedelsky method, Angoff method, Ebel method, Jaeger method, and Contrasting Groups methods) (Zieky, 2001). However, inventors of these methods did not provide specific procedural guidelines for using them. From the 1980s, the literature and practice of standard setting began to focus on how to select and use the methods; detailed descriptions of those early methods and their variations and of any new methods became readily accessible (Zieky, 2001). In addition, the literature saw a number of standard setting studies using a single method, and sometimes more than one method for method-comparison.

Despite the seeming rigor in standard setting regarding its methods and procedures, standard setting has never become free of controversies, uncertainties, and sometimes ambiguities. Fundamentally, standard setting involves judgment about what the performance standards should be and where to put the decision lines. The activity of standard setting is laden with judgments, which are based on the values and beliefs of the people who are involved. Cizek (2001) reflected after his years of work in the field that “standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other” (p. 5). Zieky (2001) concluded that the field has recognized that there are no “true” or “correct” cutscores, and cutscores are basically constructed by a group of people based on their values. Nevertheless, involving judgment does not mean that such decisions are necessarily “arbitrary” or “capricious.” As a matter of fact, all the efforts for procedural rigor in selecting and using the methods and for evaluative rigor in validating the process aim to arrive at more valid and defensible decisions in standard setting (Hambleton & Pitoniak, 2006).

Status Quo of Standard Setting in Second Language Education Literature

Although standard setting has enjoyed half a century's research and practice in general education, it hardly drew any attention from the field of second language education up to now. Among the limited literature, Brown and Hudson (2002) and Brown (2005) provided overviews of standard setting procedures when discussing test validity issues in relation to language testing situations. They particularly emphasized how standard setting relates to test consistency and test validity. They pointed out that for highly reliable or dependable tests, the accuracy of the decisions based on carefully derived cut points should be high. But for tests that are not adequately reliable or dependable, they suggested sorting out examinees who score within one standard error of measurement (SEM) or one confidence interval (CI) of the cut points, and gathering additional information for deciding on these examinees' categories. Brown and Hudson (2002) and Brown (2005) also pointed out that standards and cut points not only affect test validity in measuring what a test intends to measure, but bear upon test decision validity in interpreting test scores and using test results in the manner for which the test was designed.

Rather unfortunately, there are scarcely any standard setting studies reported in the second language education literature. The only one I was able to locate is Yoko Kozaki's (2004) standard setting study for a certification test of medical translation from Japanese to English. Kozaki used modified versions of the Angoff method, but she mainly explored the use of G-Theory and Multifaceted Rasch Analysis to check and examine the validity evidence of the cut-point decisions, and to adjust the decisions.

The Current Study

The study reported here is a pilot standard-setting study for the listening subtest on the placement test used in the English Language Institute (ELI) at the

University of Hawaii at Manoa (UHM). The ELI provides academic English courses in the skill areas of listening/speaking, reading, and writing to matriculated international and immigrant students whose TOEFL scores range between 500 and 600. For each skill area of the ELI, there are two levels of courses, i.e., intermediate and advanced. For instance, the ELI listening/speaking curriculum area has ELI 70, which is the intermediate level listening/speaking course, and ELI 80, which is the advanced level listening/speaking course. Students are required to take the ELI Placement Test (ELIPT) right before the start of the semester of their enrollment into the university. The ELIPT comprises an essay-writing test, two listening sub-tests (a dictation test and a multiple-choice academic listening test), and two reading sub-tests (a gap-filling test and a multiple-choice reading comprehension test). Based on their ELIPT results, their TOEFL scores in each skill area, and their academic background (e.g., A.A. degree earned or not; number of credit hours earned in other U.S. universities or colleges; etc.), the students are placed into intermediate or advanced levels of each ELI curriculum area, or exempted from any of them. But all undergraduate students are required to take ELI 100, which is the advanced level writing course for undergraduates, unless they have taken it in another U.S. university or college. Students who are placed into the intermediate level course of a skill area are then required to take the intermediate and the advanced level courses sequentially, after which they can be exempted from that skill area. Students who are placed into the advanced level course of a skill area will then need to take the advanced level course before they are exempted from that skill area.

The cut points for the ELIPT listening and reading sub-tests have been set at 50 (i.e., the *T* score mean for all scores) for placement into the advanced level courses, and at 60 (i.e., the *T* score one standard deviation above the mean) for exemption from the curriculum areas. Students who score below 50 in both of the ELIPT listening sub-tests and below 50 in their TOEFL listening test will be placed into ELI 70, the intermediate level course. Students who score between 50 and 60 in either of the ELIPT listening sub-tests or between 50 and 60 in their TOEFL listening test will be placed into ELI 80, the advanced level course.

Students who score above 60 in either of the ELIPT listening sub-tests or above 60 in their TOEFL listening test will be exempted from the listening/speaking curriculum area. The same placement decisions apply to the ELI reading courses. Unfortunately, the way the cut points for the listening and reading sub-tests were set was rather random and arbitrary. Basically, the cut points were originally set without using any standard setting procedures. Arguably those cut-point decisions may not be capricious in that there is consistency in placing the students with similar ability levels. But no standard setting procedure will necessarily result in arbitrary decisions, which may misplace students, do disservice to institutional functioning, and cause negative or unpleasant consequences to the students, the teachers, and the institution.

Therefore, the current study undertook the task of standard setting for the Academic Listening Test (ALT) of the ELIPT. It is a pilot study in the sense that it is the first attempt to use standard setting for any subtest of the ELIPT, and there is no previously reported similar study in the L2 literature. The researcher largely drew on the literature in mainstream education. The ALT is a listening subtest of the ELIPT, which tests students' academic listening ability in a multiple-choice format. The current version of ALT was designed by a PhD student in the Department of Second Language Studies, and has been used since Fall semester in 2005. The ALT is made up of three short lectures of about five minutes each and two longer lectures of about eight minutes each. There were a total of 35 items in the test. During the test, students listen to each lecture in order, and take notes at the same time if they wish. After each lecture, they are asked to read and answer the multiple-choice questions about the lecture. The reliability of the ALT, based on its administrations on 597 students in the past four semesters, was 0.74 (K-R20).

The main research question for the present study is thus clear and straightforward: What are the most reasonable and defensible cut-points for the Academic Listening Test of the ELI Placement Test? Three main methods were used to address those issued in this study: (a) Yes/No standard setting method, (b) Compromise standard setting method, and (c) Phi (lambda) indices at various cut points. In the following three sections, I will describe how each method was selected and used, and report the results from each of them.

YES/NO STANDARD SETTING METHOD

Selection of the Method

The yes/no method (Impara & Plake, 1997) is an alternative to the most widely used Angoff method (Angoff, 1971) for standard setting for multiple-choice tests. The Angoff method is a test-based standard setting method, which is suitable for the current study on a placement test which cannot possibly be examinee-based. In addition, the Angoff method is relatively easy to use compared to other test-based methods. As a matter of fact, Angoff (1971), right in the text, suggested using the yes/no method, which simply asks the judges or panelists to decide whether a target group of examinees can answer each item on a test correctly. In contrast, the Angoff method widely reported in the literature was actually found only in footnotes to his text. This Angoff method asks the judges or panelists to decide the proportion of a target group of examinees who can answer each item correctly. The exact procedures for using the traditional Angoff method and the yes/no method were largely developed later by other measurement professionals.

Impara and Plake (1997) used alternatives to the yes/no method in their studies by asking the judges or panelists to conceptualize a typical examinee of the target group when making their judgments, and by arguing that the aggregated judgment on the items over the judges using the yes/no method could accurately reflect the target examinee group's performances. For comparison purposes, their studies on two elementary school mathematics tests used both the yes/no method they proposed and the traditional Angoff method. The results showed that there were essentially no differences in the cut points produced by these two methods. However, the panelists reported that it was much easier to use the yes/no method than the traditional Angoff method. In addition, the results of these studies showed that there were almost no differences in the cut points produced in the first round and the second round of judgments when using the yes/no method. In the second round of judgments, the panelists were shown the normative data of the test items, and the results from the first round of judgments and its consequences.

Participants

Among the five ELI listening/speaking curriculum teachers in Spring semester 2007, three of them participated in the study as panelists. The other two were not able to come to the panel session. The researcher was one of the participants. These teachers were selected since they have the most knowledge and experience with the curriculum and the students. One of the participants has taught both ELI 70, the intermediate level course, and ELI 80, the advanced level course. The other two teachers have only taught ELI 80, but have observed ELI 70 classes and have taught students who were promoted from ELI 70.

Procedures

The procedures used for this method were largely based on the yes/no method proposed by Impara and Plake (1997). But only one round of judgments was used due to the limited amount of time that the panelists could devote to the panel session. I also referred to Hambleton (2001) when sketching out the procedures. Below are detailed descriptions of the procedures used for this method, displayed in chronological order:

1. Clarification of the purpose of the study; Brief explanation of the procedures used in the panel session. At the beginning of the session, I made the purpose of the study clear to the panelists. I explained to them the function of cut points on the ELIPT in placing ELI students, and the fact that the current cut points for the ELIPT listening subtests were set arbitrarily. I also explained how cut-point decisions would affect the students, the teachers, and the institution. I informed them that the aim of the panel session was to arrive at more valid cut points for the ALT, and that the method used came from the literature about setting cut points. In addition, I briefly explained to the panelists how the panel session would proceed. Basically, I told them that we would go though the ALT in the same way that students do, but pause in between the lectures to make judgments about which levels of students (ELI 70, ELI 80, or Exempt) would be able to answer the items correctly. I also told them that before we went through the whole test and made a judgment on each item, we would have a short training and practice session in order to make more consistent judgments.

2. Training about the distinctions among ELI 70, ELI 80, and Exempt. The purpose of the training was to help the panelists to develop a good idea of the performance standards for the three levels of placement: ELI 70, ELI 80, and Exempt. I prepared and distributed to the panelists a handout (see **APPENDIX A**) about the distinctions among the three performance levels in terms of listening abilities. I selected the best possible information about the distinctions among the three levels from the ELI official documents. Basically I emphasized to the panelists that, ELI 70 students need to develop both general and academic listening skills and need to have two semesters' listening instruction from the ELI in order to function well in their academic listening tasks; ELI 80 students need to develop academic listening skills, especially critical listening skills, and need to have one semester's listening instruction from the ELI in order to deal well with their academic listening tasks; Exempted students have already developed adequate academic listening skills and do not need any listening instruction from the ELI in order to perform well in their academic listening tasks. In addition, according to the course goals and objectives, ELI 70 students will be exposed to intermediate-level academic listening materials, whereas, ELI 80 students will be exposed to advanced-level academic listening materials. The distinctions among the three levels in terms of listening are rather generic, and no listening subskill distinctions are available for ELI 70 and ELI 80. In this way, the panelists could only have a general idea of the performance standards for the three levels of placement.

3. Training about how to use the Panel Judgment Sheet. After the training about the performance standards of the three placement levels, I distributed to the panelists the *Panel Judgment Sheet* (see **APPENDIX B**), and explained to them how they should use the sheet. I explained to the panelists that we would go through the test just as the students do, but that we would pause in between the lectures to make judgments about which levels of the students (ELI 70, ELI 80, or Exempt) would be able to answer the items correctly. For each lecture, the panelists would listen to it, take notes as they wished while listening, and then read and answer the multiple-choice questions for the lecture. They would then be provided with the answer keys for the items at the end of each lecture. By doing the test itself as the students do it, the panelists could make a better judgment in terms of which levels of students could answer the items correctly. After answering the items for each lecture and comparing their answers with the answer keys, the panelists would make a judgment for each item as to whether an ELI 70 level student, an 80 level but not a 70 level student, or only an Exempt level student could answer it correctly, and put a check in the corresponding column on the judgment sheet. It was assumed that, if an ELI 70 level student could answer an item correctly, an ELI 80 level and an Exempt student could also answer it correctly. And if an ELI 80 level student could answer an item correctly, an Exempt student could also answer it correctly. I explained to the panelists that an ELI 70, ELI 80, and Exempt level student is a typical student who would be placed into that level. I also suggested that the panelists visualize a typical student for ELI 70, for ELI 80, and for Exempt, preferably from their classes. Since the other two panelists have not taught ELI 70, I suggested that they think of a student who was currently in their 80 class, but seemed to have a lower level of listening ability and need ELI 70 listening instruction. As for Exempt, I suggested that they think of a student who was currently in their 80 class, but seemed to have developed sufficient academic listening skills and had no further need for ELI listening instruction.

4. Practice with and discussion about the first lecture of the ALT. With the *Panel Judgment Sheet* in hand, we started with the first lecture of the ALT, and used it as practice. There were a total of five questions for the first lecture. The purpose of the practice was to help the panelists check and adjust their internal judgment consistency with reference to the normative test data, and their external judgment consistency with other panelists via discussion. After the panelists listened to the first lecture and answered the questions for it, I provided them with the answer key, as well as the Item Facility (*IF*) index and Item Discrimination (*ID*) index for each of the items on a slip of paper. The *IF* and *ID* indices are based on 597 examinees' responses. I explained to the panelists what *IF* and *ID* are and what they can show about an item and students' performances on the item. I also let them know moderately acceptable ranges of *IF* and *ID* for a good item: for *IF*, 0.30- 0.70; for *ID*, above 0.30, ideally above 0.40 (Brown, 2005, p. 75). We then looked at a couple of items together and had discussions about them. For example, for one of the items that has an *IF* of 0.74 and an *ID* of 0.31, the panelists thought it was a very easy item with a rather high *IF* and a minimally acceptable *ID*, and reasoned from the lecture itself that the item was about a main point that was repeated twice at the beginning of the lecture. Thus, we thought a student placed into ELI 70 could answer this item correctly. For another item that has an *IF* of 0.28 and an *ID* of 0.35, the panelists thought it was a very difficult item with a rather low *IF* and a minimally acceptable *ID*, and reasoned from the lecture itself that the item was of a minor detail that was hard to catch. Therefore, for this item, we thought only an Exempt student could get it right. After the discussion, I asked the panelists to check over their judgment consistency based on the normative data provided, to see whether items with similar *IF* and *ID* are judged into similar placement categories. The panelists then made any changes of their judgment that they deemed necessary.

5. Panel judgment of the whole test. After practicing with and discussing the items for the first lecture, the panelists moved on to the whole test. For all the other four lectures and 30 items, we followed the procedures outlined in # 3 above. I then provided the panelists with only the answer key for each of the items, with no other normative data like *IF* and *ID*, and with no discussion among the panelists. When the panelists finished with the whole test and their judgments, I collected the judgment sheets, the test booklets, and the other materials I had distributed.

Results

Table 1 below shows the results from the panel judgment session using the Yes/No standard setting method. The three panelists were labeled as Panelist A, B, and C. For each panelist, the total number of the items for which they put a check in the 70, 80, and Exempt columns of the *Panel Judgment Sheet* was counted and listed in the 2nd, 3rd, and 4th rows of Table 1. For example, for Panelist A, there were a total of 22 items that he/she judged an ELI 70 level student could answer correctly, a total of 10 items that he/she judged an ELI 80 level but not a 70 level student could answer correctly, and a total of 3 items that he/she judged only an Exempt level student could answer correctly. Averages were then taken from the three panelists for the total number of items that they judged as ELI 70, 80, and Exempt. Table 1 shows that, on average, there were a total of 19.33 items that the panelists judged an ELI 70 level student could answer correctly, a total of 11.67 items that they judged an ELI 80 level but not a 70 level student could answer correctly, and a total of 4.00 items that they judged only an Exempt level student could answer correctly. The cut points (raw scores) derived from the averaged judgment would be 19.33 for ELI 80, and 31.00 (= 19.33 + 11.67) for Exempt. When converted to T scores, these cut points would be 49.71 for ELI 80, and 71.95 for Exempt. Based on the panel judgments, students who score below 49.71 would be at the ELI 70 level; students who score between 49.71 and 71.95 would be at the ELI 80 level, and students who score above 71.95 would be at the Exempt level. When compared with the cut points that are currently used, the cut point of 49.71 for ELI 80 is close to the currently used one, which is 50 (the *Mean*), and the cut point of 71.95 (more than two *SD* above the *Mean*) for Exempt is much greater than the currently used one, which is 60 (one *SD* above the *Mean*).

Table 1
Panel Judgments

	ELI 70	ELI 80	Exempt
Panelist A	22	10	3
Panelist B	19	10	6
Panelist C	17	15	3
M	19.33	11.67	4.00
SD	2.05	2.36	1.41
Cut Point (raw score)	19.33	31.00	
Cut Point (T score)	49.71	71.95	

In order to find out the consistency of judgment among the panelists, inter-panelist reliability was calculated and displayed in Table 2 below. Pearson product-moment correlation coefficients were used for the calculation. For each panelist's judgment on the items, ELI 70 was coded as 1, ELI 80 was coded as 2, and Exempt was coded as 3. Please note that the ratings of ELI 70, ELI 80, and Exempt are nominal data. Even though they were coded as 1, 2, and 3, the use of Pearson *r* for calculating the inter-panelist reliability violated the required assumption that the ratings should be continuous data. The researcher was not able to use any other more appropriate statistic to do the calculations for these nominal data. Thus, the inter-panelist reliability coefficients in Table 2 would not be the most accurate. From what was calculated here, the inter-panelist reliability was not very high, especially that between panelists A and C ($r = 0.37$). But since the final judgments were taken from the average of the panelists' ratings the mean (*M*) of the correlation three coefficients (using the Fisher *z* transformation) is given as well as the three-panelist reliability using the Spearman-Brown prophecy formula based on the average correlation, which turned out to be 0.76.

In addition, with a larger number of panelists, the overall inter-panelist reliability could be enhanced. In addition, the lack of clear standards may have caused difficulty in making consistent judgments.

Table 2
Inter-panelist Reliability

Panelist	A	B	C
A	1.00		
B	0.58	1.00	
C	0.37	0.58	1.00
$M_{correlation}$	0.51		
Three-panelist reliability	0.76		

COMPROMISE STANDARD SETTING METHOD

Selection of the Method

In comparison to “absolute” methods like the yes/no method used above which involve either test-based or examinee-based panel judgments, “relative” standards take into account major stakeholders’ views and perspectives about cut-point decisions especially in terms of the consequences of the decisions. These “relative” standards then involve consideration of the political nature of standard setting. Compromise methods (See Hambleton & Pitoniak, 2006 for a review of the methods) can be used in the standard-setting process to come up with “relative” standards. These standards can be combined with the results from the “absolute” methods for making more reasonable decisions.

Participants and Procedures

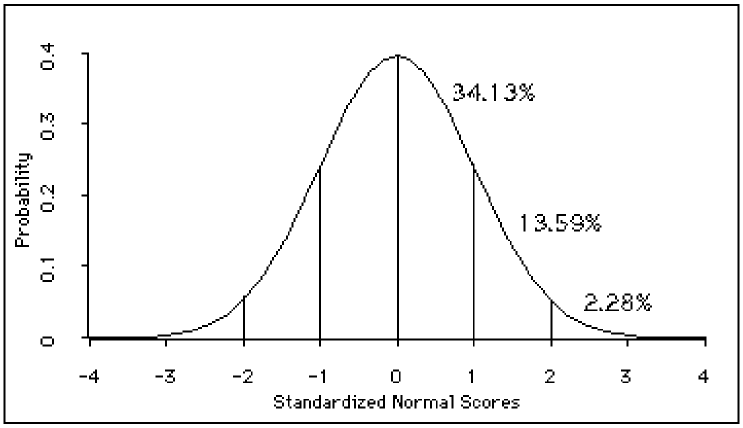
I therefore interviewed the Director and the Curriculum Coordinator of the ELI, since they are the people who supervise and take care of all aspects of the program, who have good knowledge of the curriculum and the students’ levels, and who make decisions over important aspects of the program such as cut points for the placement tests and placement policies. I planned to ask them the following three main questions:

1. Based on your knowledge about ELI students’ levels, for a single semester’s administration of the ELIPT, what percent of students do you

think should be placed into ELI 70 level (needing two semesters' instruction), ELI 80 level (needing one semester's instruction), or be exempted from listening/speaking curriculum area (needing zero semester's instruction)?

2. In relation to institutional functioning, as administrators, have you thought about what would be the ideal percentages of 70, 80, and Exempt level students, e.g., for the listening and speaking curriculum? Why so? How would the percentages (where we put the cut points) affect the number of students we have, number of sections we open for 70 and 80 level classes, number of GAs we can hire, students who get exempted from the curriculums, students' perceptions and attitudes about ELI courses, etc.?
3. The university will not provide the Asian Pacific Tuition Differential Exemption to incoming students from the Asia Pacific region from Fall 2007 semester. Considering this drop of funding for prospective international students, do you expect a drop in our enrollment? If so, how big a drop do you expect? How do you plan to deal with the situation?

In addition, after asking the first two questions, I used the figure and the table below to show and explain to the administrators the rough distribution of students placed into each level with our current cut points. I also explained to them how changes of where we put the cut scores could alter the number of students we would have for each level of placement.



Level	Percentage	If *N* = 100	If *N* = 200
70	50%	50	100
80	34.13%	34	68
Exempt	15.73%	16	32

Results

In terms of the first two questions, unfortunately, I could not get direct answers. It appeared to me that the administrators did not see the political nature of setting cut points, or think much about how manipulation of cut points could affect the institution in many ways. They thought the percentages of students placed into each level would vary semester by semester, depending on the students. They also thought that the current cut points worked quite well in placing the students accurately, since there was not much moving between the levels after the diagnostic activities at the beginning of each semester. But after asking the first two questions, I did inform them how the current cut points were placing students, and how any change in the cut points could affect the number of students we could have. Maybe this will help them to put more thought into cut-point decisions for the ELIPT in the future. As of the third question, the administrators did not yet know how the drop of funding for students from the Asia Pacific region would affect the ELI student enrollment. They would have to wait until much later to get the information.

PHI (LAMBDA) INDICES AT VARIOUS CUT POINTS

It is a fact that the consistency of decisions based on test scores may vary depending on where the cut points are set (see Brennan, 1980, 1984; Brown, 2007). Therefore, it is often useful to examine the dependability of the scores at various cut points. The phi (lambda) dependability index, or $\Phi(\lambda)$, can be used for this purpose. The $\Phi(\lambda)$ index is always the lowest at the point corresponding to the mean of the test scores, and it gets larger as the scores deviate further from the mean. This means, the dependability of decisions is always the lowest if the cut point is set at the mean of the test scores.

Table 3 below shows the phi (lambda) dependability indices at various cut points (*T* score) for the ALT. As the table shows, the dependability of the decision is the lowest at the mean (*T* score = 50), which is 0.71. This index is actually equal to the K-R21 reliability of the test. However, the current cut point between ELI 70 and 80 is right at the mean, which makes this decision point is less dependable than any other option, and the dependability of the current decisions separating ELI 70 and 80 students is 0.71, which is not ideal by any means. As for the dependability of the current decision to place Exempt students, the Φ(λ) for the cut point (*T* score = 60) is 0.85, which is better, but still not ideal.

Table 3
Phi (lambda) Dependability Indices at Various Cut Points

Cut Point (*T* Score)	**Decisions**	Φ(λ)
35		0.91
38		0.88
41		0.84
44		0.78
47		0.73
50	ELI 80	0.71
53		0.73
56		0.78
59		0.84
60	Exempt	0.85
62		0.88
65		0.91
68		0.93
71		0.95

DISCUSSION

The cut points for the ALT derived from the Yes/No method were 49.71 for ELI 80 and 71.95 for Exempt. The Compromise method used did not produce findings relevant for the cut-point decisions. The cut point of 49.71 for ELI 80 is almost the same with the currently used one, which is 50 (the *Mean*). The cut point of 71.95 for Exempt is much greater than the currently used one, which is 60 (one *SD* above the *Mean*). Based on the Yes/No method used, the current cut point for ELI 80 seems to be working well in distinguishing ELI 70 and ELI 80 students. Whereas, the current cut point for Exempt is rather problematic. For the given test, a cut point as high as 71.95 (a bit more than two *SD* above the *Mean*) may distinguish ELI 80 and Exempt level students more accurately. This means that only about 2.28% of the students taking the ALT may be qualified to be exempted. The cut point as high as this could show that the ALT does not have a sufficient number of items to discriminate well for the Exempt level students, and thus needs to be revised in this regard. The consequences of using the current ALT and the current cut point for Exempt include the possibility that we are exempting students who actually need ELI 80 instruction, and having a higher enrollment in the ELI, etc.

However, the above findings should be interpreted with caution, especially due to the small number of panelists making the judgments. A larger number of panelists could increase the validity of the findings. In addition, the Yes/No method used and the procedures involved as described above in the methods section should be carefully scrutinized and evaluated. Methods comparison studies have shown that different methods may produce different cut points. As for the procedures, there is a large number of variables that can alter the results produced. Of particular interest are variables like whether and what normative data should be provided and how much discussion the panelists should have during the process. Such variables are controversial in the literature on standard setting. Lastly, the Compromise method was not used successfully in this study. Otherwise, it could possibly provide additional useful information for making the decisions.

Although the cut point separating ELI 70 and 80 derived from the Yes/No method is virtually the same with the currently used one, which is 50 (the *Mean*), the phi (lambda) index shows that the dependability of this decision is only 0.71, not as high as might be hoped. One solution to this is to revise the test and include more items at the difficulty level of the cut points based on Item Response Theory, which will make the decisions at the cut points more dependable.

CONCLUSIONS

The study reported here aims at arriving at more reasonable and defensible cut points for the ALT of the ELIPT. As a pilot study, I think it has fulfilled its purposes to the degree it was possible. It not only provided evidence for what would be more reasonable cut points for ELI 80 and Exempt, but also offered information for test revision and development. As mentioned in the interpretation part above, the study has its limitations. In particular, the number of panelists for the Yes/No method was small, and the Comprise method was not used to its fullest extent. Perhaps the teachers should have been involved in participating in the Comprise method as well, asking for their opinions of ELIPT placement decisions and the consequences. It would also have been interesting to see how the results would differ if there were two rounds of judgment in the Yes/No method, with normative data provided for all the items in the second round of judgment and more discussions and feedback throughout the process.

In terms of the local value of the study, there is the question of whether the ELI should pursue rigorous and systematic standard setting for all the sub-tests in ELIPT. The ethical answer to the question is “yes’. Then it would require resources, particularly people, for doing the work. The administrators may have to investigate the possibilities for pulling together or bringing in resources to do this.

But one apparent problem with standard setting for the listening and reading subtests of the ELIPT is the lack of clear objectives for the performance levels. Without clearly specified objectives, the standard setting and the cut-point decisions could be ambiguous. In order to have clear objectives, the listening/speaking and the reading curriculum do at least need subskills as objectives that distinguish the curriculum levels clearly.

As for the broader value of this study for second language education, it apparently adds to the limited body of work on standard setting in the field by providing an actual standards-setting study. In second language education, placement tests are widely used. Thus, the cut-point decisions for the placement tests are quite important in affecting a large number of people's life and the language programs. At the same time, placement test have certain characteristics as a type of test, and the standard setting method and procedures for this type of test are probably unique in nature. Thus, it would be interesting and worthwhile to futher investigate how and how well standards are being set for this type of widely used test in second language programs.

REFERENCES

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Brennan, R. L. (1980). Applications of generalizability theory. In R. A. Berk (Ed.) *Criterion-referenced measurement: The state of the art* (pp. 186-232). Baltimore: Johns Hopkins.

Brennan, R. L. (1984). Estimating the dependability of scores. In R. L. Brennan (Ed.) *A guide to criterion-referenced test construction* (pp. 231-266). Baltimore: Johns Hopkins.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown, J. D. (2007). Multiple views of L1 writing score reliability. *Second Language Studies, 25*(2), 1-31.

Cizek, G. C. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. C. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Washington, DC: American Council on Education.

Impara, J. C. & Plake, B. S. (1997).Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*, 355-368.

Kozaki, K. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*(1), 1-27.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 19-51). Mahwah, NJ: Lawrence Erlbaum.

APPENDIX A

Distinctions among ELI 70, 80, and Exempt
(listening part)

ELI 70 focuses on improving listening and speaking skills, both general and academic … ELI 70 is designed for students who have less listening/speaking experience and limited familiarity with academic English and/or limited proficiency, and thus serves as a bridge to ELI 80.

ELI 80 focuses on further developing academic listening and speaking skills … This course is designed for students who have considerable listening/speaking experience and fairly advanced proficiency in English, but have only moderate familiarity with academic English and limited experience with academic listening and speaking tasks that are common in university classes.

(adapted from ELI Student Handbook: What classes are offered in the ELI?)

Major distinctive features between ELI70 and 80

Listening	**Students in 70 in general seem to lack general listening comprehension skill**, i.e., to understand what they listen to. They should first understand what they listen to (general listening comprehension skill) in order to critically respond to the listening material (critical listening skill). For this reason, the improvement of the general listening comprehension skill (or what is called “fluency in listening” in this chart) is emphasized in 70. Although it is also required **in 80, more emphasis is given on critical listening skill than on general listening comprehension skill**.

(adapted from ELI L/S Level Separation Chart (updated on April 23, 2004))

ELI 70

1. Students will develop their ability to comprehend academic listening materials[1].
• Students will learn to use pre-listening strategies (e.g., obtaining background information, having discussions to activate prior knowledge, determining contexts), **during-listening strategies (e.g., note-taking, paraphrasing, circumlocution, making inferences, predicting, getting main ideas, getting details)**, and post-listening strategies (e.g., reviewing notes, having group/class discussions) for listening comprehension of academic lectures. • Students will **become aware of** the nature of academic lectures (e.g., discourse markers used in academic lectures, emphasis of important points, use of visual aids). • Students will learn how to effectively take notes during lectures. • Students will become familiar with English pronunciation system for comprehension purposes. • Students will be exposed to **intermediate-level academic listening materials**.
Course description for ELI 70
This course provides students the opportunity to improve their academic as well as general listening and speaking skills. Particular attention is given to the comprehension of academic lectures, delivery of presentations, and participation in discussions. This course is designed as a bridge to the next level of Listening/Speaking class, ELI 80.

ELI 80

1. Students will learn to efficiently comprehend academic listening materials[1].

- Students will review pre-listening strategies (e.g., obtaining background information, having discussions to activate prior knowledge, determining contexts), **during-listening strategies (e.g., note-taking, paraphrasing, circumlocution, making inferences, predicting, getting main ideas, getting details)**, and post-listening strategies (e.g., reviewing notes, having group/class discussions) for listening comprehension of academic lectures.
- Students will be able to determine useful listening strategies that work for themselves.
- Students will **become familiar with** the nature of academic lectures (e.g., discourse markers used in academic lectures, emphasis of important points, use of visual aids).
- Students will learn how to take notes effectively during lectures.
- Students will become familiar with English pronunciation for comprehension purposes.
- Students will be exposed to **advanced-level academic listening materials**.

2. Students will learn to listen critically to academic listening materials.

- Students will learn to evaluate the contents that they comprehended.
- Students will learn to use what they just heard in order to construct their own opinions.
- Students will learn to incorporate their opinions or findings from other sources (e.g., reading materials) to respond to the listening materials in a critical manner.

Course description for ELI 80

This course provides the students with the opportunity to further improve their academic listening and speaking skills to enable the students to follow lectures and participate orally in class in an American university setting. The course will focus on listening comprehension, presentation, and discussion skills. This course is designed for students who have considerable listening/speaking experience and advanced proficiency in English as an additional language.

(adapted from Goals and Objectives—ELI Listening & Speaking—(Updated on November 20, 2003))

APPENDIX B
Panel Judgment Sheet

Panelist Name:____________

Method: Adaptation of Yes/No Method (Impara & Plake, 1997)
Directions:

- Put a √ in the 70 level column, if you think a 70 level student can answer the item correctly.
- Put a √ in the 80 level column, if you think an 80 level student, but not a 70 level student, can answer the item correctly.
- Put a √ in the Exempt level column, if you think only an Exempt level student can answer the item correctly.

Item	Your Answer	70	80	Exempt
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				

Directions:

- Put a √ in the 70 level column, if you think a 70 level student can answer the item correctly.
- Put a √ in the 80 level column, if you think an 80 level student, but not a 70 level student, can answer the item correctly.
- Put a √ in the Exempt level column, if you think only an Exempt level student can answer the item correctly.

Item	**Your Answer**	**70**	**80**	**Exempt**
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				