Research Forum

Score Reliability and Placement Testing

Paul A. Westrick
Kyushu University

This study examines the piloting of a commercially-produced test of English, the Quick Placement Test – Pen and Paper Test (QPT-PPT). In consecutive administrations of two versions of the test with 161 first-year students at a Japanese university, the test results failed to discriminate among students of varying proficiencies. Narrow ranges, low score reliability estimates, and large standard errors of measurement characterized the results. Item analysis revealed that most of the test items did little to separate high and low scoring students. The data also suggests that test anxiety, familiarity with the test format, and test-taking skills were important factors in the test scores.

本論文はQuick Placement Test – Pen and Paper Test (QPT-PPT)というクラス編成テストの妥当性検証の報告である。総計161名の日本の大学生に2つの版のQPT-PPTを行った結果、このテストはさまざまな能力の学生を弁別することができないということがわかった。得点分布の範囲は狭く、信頼性も低く、標準誤差も非常に低かった。また項目分析の結果の示すところ、ほとんどの項目が下位の受験者と上位の受験者を弁別することができていなかった。さらに、テスト不安、テスト形式、受験技能なども大きく結果に影響していることも示唆された。

Throughout Japan, university and college teachers of English frequently complain that their institutions do not place students in classes based on English proficiency levels. Because creating a placement test requires some knowledge of testing principles and can be time consuming, purchasing a commercially-produced proficiency test may appear to be an easy solution. On the other hand, commercially-produced tests are often expensive, and scoring them can take time. For these reasons, it was decided to independently pilot test the Quick Placement Test – Pen and Paper Test (QPT-PPT), an inexpensive commercially-produced placement test, created by the University of Cambridge Local Examinations Syndicate (UCLES) and published by Oxford University Press (OUP).

Research Focus

Placement Tests

There are a variety of reasons why teachers want placement systems at their institutions. One is that mixed ability classes are hard to teach effectively. Without a placement system, teachers typically struggle to aim instruction in ways that do not simultaneously befuddle lower-level students and bore higher-level students. In an attempt to reach all students at all ability levels, teachers may try to create different activities for different groups within a single classroom, "mini-lessons" that try to please everyone but often please no one. In contrast to this inefficiency, Redondo (2000) points out that, "Teachers can use materials, as well as teaching styles and pedagogical approaches more effectively when classes share a greater similarity of learning qualities and characteristics" (p. 126). Besides classroom teaching efficiency, another problem of mixed-level classes is grading. Teachers often face the dilemma of whether to lower standards to pass hard-working low-level students or to maintain standards and fail low-level students who try their best but simply cannot compete with their classmates. Unless the instruction is concentrated on their needs, the weakest students invariably fall further behind their peers. Another issue intertwined with the others, and probably the most important, is the amount of learning that takes place. As Chauncey and Frederiksen (1951) recognized long ago, for a student to advance, that student should "take courses which are neither too difficult nor which involve wasteful duplication of earlier-learned content" (pp. 108–109). Because of the problems associated with mixed ability classes, many teachers, as Brown (1996) observes, tend to prefer having students grouped by ability levels, bringing together those with similar strengths and weaknesses. With such a class, the teacher can concentrate on addressing their weaknesses in ways that are appropriate for those students.

These teachers generally want students to be sorted with a placement test, and here it is necessary to differentiate between placement tests and proficiency tests. Though they are similar in that they are normreferenced tests (NRTs), in which students' performance is compared to that of other test takers, they are not identical, as Brown (1996, pp. 11–12) makes clear:

> Examining the similarities and differences between proficiency and placement testing will help to clarify the role of placement tests. To begin with, a proficiency test and a placement test may look very similar because they are both testing fairly general material. However, a proficiency test tends to be very, very general in character, because it is designed to assess extremely wide bands of abilities. In contrast, a placement test must be more specifically related to a given program, particularly in terms of the relatively narrow range of abilities assessed and the content of the curriculum, so that it efficiently separates the students into level groupings within that program.

Hughes (2003) expresses similar thoughts and advises that placement tests should be developed by the users themselves so that they specifically meet their needs. As for commercially-produced tests, he states that:

> Placement tests can be bought, but this is to be recommended only when the institution concerned is sure that the test being considered suits its particular teaching programme. No one placement test will work for every institution, and the initial assumption about any test that is commercially available must be that it will not work well. (p. 16)

Ideally, an institution's placement test should be connected with its curriculum, but institutions still choose commercially-produced proficiency tests as placement tests for a variety of reasons, some more defensible than others. Starting with the least defensible, one reason given for choosing a commercially-produced test is status. Often administrators, and unfortunately some language teachers, believe that they can improve the image of their school by using a famous test or a test produced by a famous institution. Another poor reason (or excuse) given is that creating a placement test is too difficult. A lack of enthusiasm or confidence often leads to the decision that it is best to defer to the experts, the producers of commercial tests. At some institutions, a dedicated group of teachers may win agreement or approval for a placement test, but are stifled in their struggle for the development of a true curriculum. "Curricula" that consist of little more than course titles are, sad to say, not uncommon. Without institutional guidance, instructors teach whatever they find appropriate with little or no regard for course titles or what other instructors teach. In such a situation, a test that emphasizes specific language targets may be no better than a commercially-produced general proficiency test. Probably the most defensible reason that can be given for using a commercially-produced test is that such a test is used only as a temporary measure until it can be replaced with a customized test that matches a true curriculum, a curriculum that has clearly defined goals and objectives (for a more detailed discussion of testing and curriculum development, see Brown, 1995, 1996).

For those institutions that decide to implement a placement system, one last critical factor is time. As noted in Poel and Weatherly (1997), and Fulcher (1997), the brief time between when students arrive and when classes begin often poses a problem for placement personnel, and the amount of time allotted for placement testing may be less than optimal. Ideally, the placement test should be quick and easy both to administer and to score. The initial attraction of the QPT-PPT is that it can be administered in 30 minutes, and the answer sheets can be scored on campus with transparent overlays. The trade off is that with a commercial test there is the possibility that many of the items are not written at an appropriate level for the examinees at a given institution. In Classical Testing Theory (CTT), generally speaking, longer tests produce scores that are more reliable than short tests (Brown, 1989; 1996). In Item Response Theory, test length may not be a critical factor, but even in CTT it is recognized that if a short test contains only good items that clearly separate the examinees, the score reliability estimate may be high. But if a short test is filled with items so easy that almost all the students can answer them correctly or items so difficult that almost all the students answer them incorrectly, there may not be enough remaining high-quality items to separate the students and reliability will suffer.

Item Analysis

Item quality has an effect on the amount of score variance, and the amount of score variance influences the reliability of the scores. If a placement test contains items that all examinees answered correctly or items that all examinees answered incorrectly, those items can be removed from the test without altering the distribution of scores or the ranking of examinees; in other words, the amount of variance would remain the same. Such items fail to separate examinees and provide no useful information for placement decisions; therefore, they can be eliminated.

Instances where all examinees answer an item correctly or all answer an item incorrectly are rare, but at what point do we decide that an item is functioning well, that is, separating students and creating score variance? Item analysis is not an exact science, but over the years testers have created some general guidelines for analyzing the quality of items. One tool test writers have is the item facility (IF) index, the proportion of examinees answering an item correctly. Brown (1996) states that IFs between .30 and .70 are generally acceptable for an NRT, and McNamara (2000) states that testers should be content with IFs between .33 and .66. Another evaluation tool is the item discrimination (ID) index, the difference between the IF for the highest-scoring third of the examinees and the IF for the lowest-scoring third of the examinees (ID = IF_{upper} − IF_{lower}). According to Ebel's guidelines (1979, cited in Brown, 1996), an item with an ID at or above .40 is considered very good. Those from .30 to 39 are considered reasonably good, those between .20 and .29 marginal, and items in both categories may be improved. Items below .19 are considered poor, and may be either rewritten for improvement or discarded. In sum, a test needs items that are neither too difficult nor too easy, and the items should discriminate between the higher- and lower-scoring students. Such items are needed to generate score variance, which in turn affects score reliability.

Score Reliability and Validity

Reliability, the precision and consistency of measurements, and validity, measuring what one claims to be measuring, are vitally important. A scale that says a person weighs 50 kilos one minute and 100 kilos the next is not giving reliable measurements. Even if a scale does give reliable measurements, those measurements are only valid measurements of the person's weight; they would be invalid measurements of the person's height.

Discussing reliability and validity in detail is beyond the scope of this paper, but teachers should keep a few points in mind. First, a test cannot be considered reliable or unreliable because reliability is a characteristic of scores, not of the measure. For this reason, Thompson (2003a) stresses that testers and researchers should be clear by talking about the reliability of *scores*, not the reliability of tests. Two very different groups may produce widely different scores on the same measure, as Brown (1989) demonstrated with a cloze test. The second point to remember is that, generally speaking, heterogeneous groups tend to have more variance in their observed scores and therefore produce more reliable score estimates than homogeneous groups do. In the words of Thompson (2003b, p. 93, emphasis in original):

> Reliability is driven by variance – typically, greater score variance leads to greater score reliability, and so more *heterogeneous* samples often lead to more *variable* scores, and thus to higher reliability. Therefore, the same measure, when administered to more heterogeneous or to more homogeneous sets of subjects, will yield scores with differing reliability.

Third, while reliability estimates may change from group to group, there are standards to be met. This point is especially true with high-stakes tests, and as there are times when placement tests are high-stakes tests, these standards should be met. For important tests, score reliability coefficients of .90 are considered the minimum and reliability coefficients of .95 or higher are preferred (Hopkins, Stanley, & Hopkins, 1990; Nunnally & Bernstein, 1994). Considering that the highest possible reliability estimate for scores is 1.0, having a reliability coefficient of .90 appears to be quite good, but even with a reliability coefficient of .90, the standard error of measurement will be nearly one-third the size of the standard deviation (Nunnally & Bernstein, 1994).

The fourth point is that there can be no score validity without score reliability, for it is difficult to make sound judgments when the measurements of what is claimed to be measured are unreliable. As Thompson (2003a, p. 6) explains,

> Perfectly unreliable scores measure nothing. If the scores purport to measure something/anything (e.g., intelligence, self-concept), and the scores measure nothing, the scores (and inferences from them) cannot be valid. Scores can't both measure nothing and measure something. The only time that perfectly unreliable scores could conceivably be valid is if someone was designing a test intended consistently to measure nothing.

While zero reliability is the extreme, we know that we want a high reliability estimate and a small standard error of measurement (SEM), but do most teachers really understand why? Teachers generally understand percentiles and have seen distribution curves, but understanding band scores and the percentages that fall within a standard deviation (SD) in a normal distribution (a bell curve) can clarify the importance of high reliability.

In CTT, the SEM is used with an examinee's observed score to create band scores (Bachman, 1990; Hughes, 2003) to estimate the examinee's true score. We determine the SEM by multiplying the SD by the square root of one minus the reliability estimate (rxx'). That is, SEM = SD $\sqrt{1- rxx'}$. The higher the reliability estimate is, the smaller the standard error of measurement. Using one SEM above and below an examinee's observed score, we create a band score, and we can estimate with about 68 percent certainty that an examinee's true score falls within that band. Using two SEMs, we can estimate a true score with about 95 percent certainty, and with three SEMs we can estimate a true score with about 99 percent certainty. Turning now to the normal distribution, the bell curve, we know that roughly 68 percent of the people will fall within one standard deviation from the mean – 34 percent above and 34 percent below. A student with an observed score at the mean is at the 50th percentile; a student one SD below the mean is at the 16th percentile; and a student one SD above the mean is at the 84th percentile.

Now let's put this together. For example, we have a 60-item placement test, and we know that the score reliability estimate is .90; the average score is 30; the SD is 10 points; the SEM is 3.16; and the distribution is normal. A student with an observed score of 30 is at the mean, thus at the 50th percentile. Under such circumstances, using band scores that extend out three SEMs, we can estimate with 99% confidence that the student's true score is in a band between 20.52 and 39.48, a band that covers nearly two standard deviations. That band does not quite stretch all the way to the 16th and 84th percentiles, but it does come rather close. If making fair and defensible decisions with score reliability estimates at .90 is difficult, making fair and defensible decisions with score reliability estimates below .90 is more difficult, if not impossible.

Past Studies of Placement Tests

There have been a number of studies of placement tests (Blais & Laurier, 1995; Fulcher, 1997; Poel & Weatherly, 1997; Wall, Clapham, & Alderson, 1994), but most of these studies concern customized tests developed at the institutions where they were used. Studies of commercially-produced tests have been harder to find. In one study, Culligan and Gorsuch (1999) explored the use of a commercially-produced proficiency test (the Secondary Level English Proficiency test or SLEP test) as a placement test at a Japanese university and determined that it did not work very well for them. In their analysis of the test results, they discovered that most of the test items did not discriminate between the high- and low-scoring students, and the test had a high SEM. Furthermore, and more importantly, the test did not match well with their curriculum.

Although that study suggested that a commercially-produced test was inappropriate for an institution with a defined curriculum, the use of such a test at an institution with a less clearly defined curriculum could possibly be worthwhile. In a very small piloting of the QPT-PPT for review purposes at a Japanese university with liberal admission standards, the test separated students much as the instructor expected (Westrick, 2002). In two classes with a total of 27 students, the scores of most Japanese students were clustered at the low end of the scale, just above the few Chinese students who had had no formal English education, and below the Chinese and Korean students who had had formal English instruction in their home countries. Though it was not covered in the published review, removing the Chinese and Korean students decreased the range of observed scores and the score variance. The narrow range of scores for the Japanese students who made up the majority of test-takers warranted a larger piloting of the QPT-PPT.

Research Questions

The primary objective of the current study was to test the test. Ideally, the results of the administrations of the QPT-PPT would be analyzed with the students' scores from the Center Test and university entrance exam, but this was not possible. All information regarding student scores on the Center Test and the university's entrance exam is considered private and was unavailable for the study. Operating under this constraint, the research questions were as follows:

1. Would the QPT-PPT separate university students who have already been sorted by the Center Test and a university entrance exam?
2. As placement tests have serious implications for students, would the reliability coefficients of the test scores for these students be at or above 0.90?
3. As there are "two parallel (photocopiable) versions" (QPT-PPT user manual, p. 2) of the QPT-PPT, how high would the correlation between student scores on Version 1 and Version 2 be?

Information regarding the cutoff points for those tests would have been useful in regard to the first research question, and a study that correlated the students' QPT-PPT scores with their Center Test and/or entrance exam scores would have been ideal for establishing parallel-forms reliability in the third question, but as stated earlier, the students' Center Test and entrance exam scores were unavailable due to privacy concerns.

Method

Participants

With one exception, all of the participants in the piloting of the test were first-year technology or economics students at a national university (one law major was enrolled with the economics majors). They were in three different class sections: one was for economics majors, and two separate sections were for technology majors. Section sizes ranged from 53 to 60 students, and 161 of the 167 students enrolled in the first-year English classes took the tests. Of those students, 159 were from Japan, one was from the People's Republic of China, and one was from South Korea. New students are admitted to their department within the university largely based upon their performance on the national Center Test and the university's entrance examination. Once admitted to their department, students are placed in *jun* order (an "alphabetical" order based on the pronunciation of their family names in the Japanese phonetic system), assigned a student number, and then placed into sections within their department in this alphabetical/numerical order. All students within each section are required to take a general English course with the other students in their section regardless of their scores on the Center Test and the university's entrance examination. There is no initial placement system.

Materials

Created by UCLES and published by OUP, the QPT-PPT is, at first glance, quite easy to use. It is legally photocopiable, so an unlimited number of tests can be reproduced, and it can be scored locally with transparent overlays that come with the test package. The QPT-PPT tests reading, grammar and vocabulary skills only. The user manual suggests that the QPT not be used as the sole instrument to evaluate and place students, that speaking and writing skills be assessed, and that with the PPT, a listening component should be added (the QPT-Computer Based

Test, the CBT, has listening items built into the test). UCLES does not assert that the QPT is the only thing teachers need in order to make placement decisions; it is but one resource that English language programs can utilize when making placement decisions.

There are two versions of the test, Version 1 and Version 2, and each version has two parts. Part 1 has 40 multiple-choice items, and all test-takers must complete this portion of the test. (Examples of items similar to those found on the actual tests can be seen in Appendix A.) Items 1 through 5 provide test-takers with notices (signs, postings, etc.) and ask where these notices would be seen. Following the notices are three cloze passages with five items each (items 6–20); test-takers choose the best word to fill each space in the passage. Items 21 through 40 are individual sentences with a word or short phrase missing; students choose the best word or phrase to complete each sentence. For items 1 through 10, there are three possible answers for each item, and for items 11 through 40 there are four possible answers for each item. Part 2 of the test consists of an additional 20 multiple-choice items that are designed for more advanced learners. It starts with two cloze passages with five missing words or phrases for each passage (items 41–50); followed by individual sentences with a word or short phrase missing (items 51–60). All 20 items in Part 2 have four possible answers. It should be noted that at the beginning of Part 2 there are written instructions for students not to do Part 2 unless told to do so (which is important because it caused problems in the first administration in this study).

Test administrators have the option of giving their students only Part 1 of the test if they believe that their students are below the advanced level according to the Association of Language Testers in Europe (ALTE) descriptions of users' language abilities, which are provided in the QPT-PPT user manual. For administrators unsure of their students' levels, the user manual advises the use of both parts of the test (all 60 items). Test scores from the 40 item and 60 item tests can be matched to the level chart in the user manual. For the 60 item test, students with scores between 0 and 17 are at the beginner level; those between 18 and 29 are at the elementary level; those between 30 and 39 are at the lower-intermediate level; those between 40 and 47 are at the upper-intermediate level; those between 48 and 54 are at the advanced level; and those between 55 and 60 are at the very advanced level. Again, the user manual suggests that the QPT be used with other forms of assessment to help make decisions about students who score within one SEM (±4 on the 60 item test, ±3 on the 40 item test) of a cutoff point on the ALTE scale.

It should be noted that at the time of this writing, the QPT-PPT packet with answers could be purchased from booksellers on the Internet, making security a concern. The QPT-CBT may be safer because the computer checks item responses without giving the answers, but the possibility remains that someone could purchase a CBT packet and record the questions while taking the test multiple times. In this study, it appears that none of the students had access to copies of the tests prior to the day of the administrations.

Procedure

The participants took the QPT-PPT in the first semester of their first academic year. Each class section was told in English that they would take two versions of the test, Version 1 and Version 2, and they were told to do both Part 1 and Part 2 of each test. They were told that half of the class, Group 1, would take Version 1 first, and the other half, Group 2, would take Version 2 first, and that they would have 30 minutes to complete the test (This first test administration is hereinafter referred to as Administration A). After the first test was completed, the tests were collected and the second test was given. Students in Group 1, those who had taken Version 1 first, then took Version 2, and students in Group 2, those who had taken Version 2 first, then took Version 1 (This second test administration is hereinafter referred to as Administration B). They again had 30 minutes to complete the test. These instructions were also written on the chalkboard in English. Students were told to mark their answers on the answer sheet, and the instructor modeled on the chalkboard how to mark answers on the answer sheet. Students were asked if they had any questions. There were no questions.

During Administration A, the instructor walked about the classroom observing the students as they took the tests. In each of the three class sections, the instructor observed that some students stopped after completing Part 1 of their tests. When this was observed, the instructor told those students to continue on to Part 2, and a general reminder was made to the entire class that they were to do Part 2 after finishing Part 1. After thirty minutes, the tests and answer sheets were collected. After all tests and answer sheets were accounted for, students received the second test (Version 2 for Group 1 and Version 1 for Group 2) and a new answer sheet. After the second test administration, Administration B, was finished, tests and answer sheets were again collected. When tests and answer sheets had all been accounted for, the students were released.

Total administration time varied between 80 and 85 minutes for each class section. Answer sheets were hand scored using the transparent overlays provided in the test packet. (This took one marker much longer than the time needed to administer the tests. Scoring the tests with an optical scanning machine would have made this task much quicker and easier.) Scores were double checked when each student's answers were entered into a spreadsheet for analysis.[1]

Results

The data gathered from the back-to-back test administrations revealed interesting changes in how 64 students (nearly 40% of the participants) approached the test during the second administration. In Administration A, four students did not do Part 2 after completing Part 1 despite each class being told repeatedly to do so; two students apparently started with Part 2 of their tests, struggled, and were then unable to complete Part 1. It appears that these six students did not understand the teacher's instructions. Additionally, 58 students did not attempt to answer all the items on their tests. They apparently were not as test-wise as the other 97 students, who either paced themselves so as to have time to read every item or realized there was no harm in guessing on difficult items and continuing onward through the test. In Administration B, these differences largely disappeared. All students did Part 1 then Part 2 of the test, and all but 13 students attempted to answer all 60 items. Consequently, the descriptive statistics for Administration B are quite different from those for Administration A.

In Table 1, in Administration A, the students' means and medians for Version 1 (AV1) and Version 2 (AV2) were very close. Unfortunately, the ranges of scores on both tests were below 30, meaning more than half of the scale went unused. The reliability of test scores for Version 1 was 0.61 while the reliability of test scores for Version 2 was slightly lower at 0.59. For the test scores on both versions, the SEMs were more than half the size of the SDs. It also appeared that students scored better on Version 2 than on Version 1.

In Administration B, it appears that all the students started with Part 1 of their tests and attempted Part 2 after completing Part 1, that 148 of 161 provided answers for every test item, and that the 13 who did not answer every item came very close to doing so. As a result, the variance in test scores decreased. Score ranges became even narrower (20 for Version 1 and 18 for Version 2); the reliability coefficients for the test scores fell to

Table 1. Descriptive Statistics by Group and Test Version

	G1			G2			G1+G2		
	AV1	BV2	V1&2	AV2	BV1	V1&2	V1A&B	V2A&B	V1&V2
N	82.00	82.00	82.00	79.00	79.00	79.00	161.00	161.00	161.00
k	60.00	60.00	120.00	60.00	60.00	120.00	60.00	60.00	120.00
Mean	35.13	37.80	72.94	35.68	36.32	72.00	35.71	36.76	72.48
Median	35.00	37.00	72.00	35.00	37.00	72.00	36.00	37.00	72.00
Mode	36.00	37.00	71.00	35.00	37.00	67.00	37.00	39.00	71.00
Midpoint	30.50	38.50	73.00	34.50	36.50	68.50	31.50	34.50	71.50
High	44.00	47.00	90.00	47.00	46.00	86.00	46.00	47.00	90.00
Low	17.00	30.00	56.00	22.00	27.00	51.00	17.00	22.00	53.00
Range	28.00	18.00	35.00	26.00	20.00	36.00	30.00	26.00	38.00
SD	5.04	3.87	7.36	4.61	4.28	7.67	4.70	4.37	7.51
K-R20	0.61	0.41	0.65	0.59	0.46	0.68	0.55	0.53	0.66
SEM	3.13	2.98	4.33	2.97	3.15	4.33	3.16	2.99	4.36
Correlation V1V2			0.35			0.49			0.37

Note. G = Group; A = Administration A; B = Administration B; V = Version

0.46 for G2-BV1 and 0.41 for G1-BV2; and the gaps between the SDs and SEMs narrowed even more dramatically. It again appeared that students scored better on Version 2 than on Version 1, and the average scores on both versions of the test rose in the second administration.

In both groups shown in Table 1, students' scores on both versions are combined. This spread students out further than just using one version of the test. Despite the wider ranges of scores, the ranges were still narrow in relation to the scale, and the widest range of scores was 36 (Group 2), meaning only 30 percent of the scale was utilized. With regard to score reliability, the estimates for the combined scores were higher than those for a single administration, as would be expected, but for neither group did the reliability estimates approach .90.

The item analysis summaries in Tables 2 and 3 help explain the low reliability estimates. Very few test items were at the appropriate level for the students and not many items separated high and low scorers. Fewer than half of the items on both versions of the test had IFs between .30 and .70. That is, more than half of the items on both versions of the test appear to have been too easy or too difficult for the students. Furthermore, even fewer items had IDs of .40 or higher. The vast majority of IDs fell below .19, and in all but one instance there were more *negative* IDs than there were IDs above .40.

Table 2. Number of Items Meeting IF and ID Guidelines

	G1			G2			G1+G2		
Number of items ...	V1	V2	V1+V2	V1	V2	V1+V2	V1	V2	V1+V2
with IFs between .30 and .70	27	21	48	25	19	44	28	18	46
with IDs of .40 or higher	6	3	9	5	2	7	3	1	4
that met the guidelines for both ID and IF	5	2	7	5	2	7	3	1	4
k	60	60	120	60	60	120	60	60	120

Table 3. Number of Items by ID

	G1			G2			G1+G2		
ID	V1	V2	V1+V2	V1	V2	V1+V2	V1	V2	V1+V2
.40 +	6	3	9	5	2	7	3	1	4
.30 to .39	8	4	12	6	5	11	6	6	12
.20 to .29	8	10	18	7	12	19	14	12	26
.10 to .19	14	13	27	16	16	32	9	12	21
.00 to .09	20	20	40	18	21	39	22	24	46
Negatives	4	10	14	8	4	12	6	5	11
k	60	60	120	60	60	120	60	60	120

Two other observations about the individual test administrations should be made. First, the correlations (Pearson product moment correlations) between the two versions of the test were low for both groups even though they are supposed to be parallel forms, and the order in which the tests were taken seems to be a factor. Second, there may have been regression toward the mean in Administration B. Predicting the amount of regression depends on the reliability of the test scores (Hopkins, et al., 1990). With perfect score reliability ($r_{xx'} = 1.0$) there should be no regression; with perfectly unreliable scores, all examinees should be expected to regress to the mean. With low score reliability in Administration A, some regression toward the mean in Administration B should have been anticipated. It seems that regression to the mean did occur, as scores in Administration B were more tightly clustered around the mean than in Administration A. However, a closer look was warranted because of the differences in approaches in Administration A, so students were divided by their approaches to see how much their scores differed between the two administrations (see Table 4). Removing confusion and adding experience clearly made a difference.

Table 4. Changes in Total Scores Based on Approach

Students in Administration A who ...	N	Mean score Administration A	Mean score Administration B
answered all 60 items	97	36.70	37.12
did Parts 1 and 2 but did not answer all 60 items	58	34.05	36.81
did Part 1 but not Part 2	4	30.50	39.25
did Part 2 but did not finish Part 1	2	23.00	39.00

Note: The 58 students who did not answer all the items in Administration A answered an average of 8.14 more items in Administration B.

Aside from the changes in scores, the differences in how the students approached the test revealed additional information concerning reliability, correlation, and regression to the mean. Regarding reliability, removing the six who were confused by the instructions (one from Group 1 and five from Group 2) lowered the reliability estimates for G1-AV1 0.61 to 0.55 and for G2-AV2 from 0.59 to 0.55, and it decreased the ranges from 28 to 20 and from 26 to 20 respectively. Their removal improved the correlation coefficients for the two groups, from 0.35 to 0.40 for Group 1 and from 0.49 to 0.54 for Group 2, but the correlations were still low. Even when looking at only the 97 students who approached the test consistently answering all 60 items in both administrations, the score correlations were 0.48 for Group 1 (N=52) and 0.65 for Group 2 (N=45). Regarding regression toward the mean, practice and experience (Brown, 1988) could explain the improvements of 64 students and the decrease in test score variance, yet for the 97 “consistent” students, the SD for Group 1 (N=52) fell from 4.36 in Administration A to 3.71 in Administration B, and for Group 2 (N=45) it fell from 4.97 in Administration A to 4.45 in Administration B, suggesting that regression to the mean did occur in Administration B.

Discussion

Though not part of the initial research focus, test anxiety and the practice effect must be addressed. Regression toward the mean can partially explain the differences in the descriptive statistics for the first and second administrations of the QPT-PPT, particularly the decline in the reliability estimates, but the increase in the number of items answered and the slight rise in the means for both groups in both tables suggest that the students were unsure of how to approach this test the first time they took it. As mentioned earlier, a review of the data revealed that six students apparently misunderstood the directions, and 58 students apparently were not as test-wise as other students were. With experience and practice, it appears that the students became more comfortable with the test format, and the average scores for these students rose in Administration B.

Without practice and experience, many students (if not most) have some degree of anxiety before and during a test administration. For those familiar with English education and testing in Japan, it is common to see students initially paralyzed when they encounter their first "English-only" test. The students in this study had taken two English-only listening tests already—they had taken a diagnostic test on the first day of the semester, and they had taken their first test of record five weeks later. Many were visibly distressed when they took the diagnostic test, but they were much more relaxed and confident when they took their first test of record five weeks later. It seems that they made the same adjustment with the QPT-PPT within a much shorter time frame. When the students learned on the day of the test administrations that the tests were made by the University of Cambridge Local Examinations Syndicate and published by Oxford University Press, one could not miss the visible and audible responses of anxiety. The test had a high degree of face validity with the students, and many were manifestly intimidated, but their anxiety levels declined as they progressed through the tests, particularly during the second administration.

Anxiety may also have been a factor in the apparent confusion six students had with the instructions. How often have teachers looked at a student's test, seen that the student did not understand the instructions and performed poorly as a result, and then said something along the lines of, "Well, that tells me something." Yes, it tells the teacher something—the student did not understand the instructions, and as Japanese students, generally speaking, do not ask questions because they do not want to appear foolish, such a student would not ask for clarification. It does not necessarily tell the teacher that the student is significantly less proficient than the other students in whatever intended construct the teacher is trying to measure. As a case in point, the "worst" student in Administration A transformed into an "above average" student in Administration B. He gained 22 points not because his English skills improved in under an hour; it is just that with a bit of experience, he learned how to take the test. That student became test wise, as did many others in this study.

Using the data from Administration A to make judgments on English reading, grammar, and vocabulary proficiency is probably inappropriate. Anxiety, confusion, and lack of experience created variance in the scores, but it was error variance because it was "not related to the purpose of the test" (Brown, 1996, p.188). Reducing this error variance caused the range and reliability of the scores to fall in Administration B. The data collected from Administration B, in which all the students understood the test format, had prior experience with the test, and were relaxed, probably provides a more trustworthy measurement of the intended construct. One could argue that another option would be to combine the scores from both administrations, mitigating the negative effects of the first administration for some students without throwing away the other students' time and effort, but this would still be unfair to the 64 students who were not as test wise as their peers.

The disappointing results from this pilot testing do not necessarily mean that the QPT is a bad test. Again, the expression "the reliability of the test" should be avoided; "the reliability of the test scores" should be used. Closely linked to this was the fact that a test may produce reliable scores with one group of test participants but produce unreliable scores with a different group of test participants. A placement test should separate students along a continuum, and the range of scores should be wide enough to make distinctions, but with this population of students, the QPT-PPT did not separate students very well, and the reliability estimates of the scores were quite low. The QPT-PPT may very well distinguish among students who have widely different educational backgrounds. In the QPT-PPT user manual (UCLES, 2001, p. 14), it says that, "prior to publication the QPT was validated in 20 countries by more than 5,000 students." At the end of the description of the two validation phases of the QPT, the manual concludes with this sentence: "By investigating the reliability of the test scores as well as the tests themselves, we have produced a test which is both reliable and practical" (p. 14). Considering the diversity of the participants in the development of the test, it would come as no surprise that students' test scores were broadly dispersed along the continuum and that the reliability of the test scores for both versions was acceptable. (Though SEMs are given in the user manual, reliability coefficients are not given.)

EFL situations tend to be quite different from each other, as this piloting of the QPT-PPT with the first-year university students suggests. Broadly speaking, this pilot study supports Culligan and Gorsuch's (1999) observation that there is typically not much diversity as measured by global proficiency tests among the incoming student populations at Japanese universities. In Japan, students follow a national curriculum through junior and senior high school, and for those going on to university, many are then sorted first by a nationwide test (the Center Test) and again by university entrance exams. If anything, the QPT-PPT simply confirmed that students selected for admission had been screened by the Center Test and possibly the university entrance exam with regard to their proficiency in reading, grammar, and vocabulary. In this study, an additional general test of English reading, grammar, and vocabulary, with low score reliability estimates and SEMs approaching the size of the SDs, provided little information that could be used for placement decisions.

This leads to the final point made in the section on score reliability, which is that without score reliability, score validity is cast into doubt: if the measurements cannot be trusted, sound inferences cannot be made based upon those measurements. Using the data for all students in Table 1, one can see that a student with an observed score of 37 on test Version 1, Administration B, would be at the 50th percentile. Going out three SEMs ±, one can estimate with 99% confidence that the student's true score is between 28 and 46. Looking at the observed scores, which ranged from 27 to 46, and then looking at the student's band score, one could estimate that the student could be somewhere between the 2nd and 99th percentiles. Again using the data from for all students in Table 1, one can estimate that a student who scored 38 on test Version 2, Administration B, is just slightly above the 50th percentile. Going out three SEMs ± one can estimate with 99% confidence that the student's true score is between 29 and 47. This student's band score is actually larger than the range of observed scores. Observed scores of 30 and 47 put students at the 1st and 99th percentiles respectively. There is a tremendous amount of overlap when using band scores at the 99% confidence level, and though some distinctions can be made at the extremes, the "average" students in either group may very well be among the strongest or weakest students in their respective groups. Learning that a student is between the 1st and 99th percentiles after the test is not much better than knowing the student stood between the 1st and 99th percentiles before the test. Though these scores are not perfectly unreliable, they are unreliable enough to have come close to measuring nothing.

Conclusion

While score reliability matters, it is a characteristic of the group, not the test. Though the QPT may have been validated in 20 countries with over 5,000 students, the 161 students in this study came from only three countries, and 159 of them came from just one country, Japan. Other commercially-produced tests that focus on reading, grammar, and vocabulary that were validated worldwide would probably produce no better results with this population. The students in this study had been sorted by their performances on the Center Test and the university entrance exam, but their scores on these two tests had no impact on their class assignments. There may be differences between these students in regard to their English language abilities, but administering the QPT-PPT did not generate data that could be used to sort these students in a fair manner. It would probably be much better for these students to be separated by their abilities in other, as yet untested, skills. Better yet, the skills tested should be linked to the goals and objectives of the curriculum.

Sadly, there are teachers and administrators who never question the reliability of test scores and instead accept raw and converted scores as perfect reflections of students' abilities. Cut points are set and strictly followed, and decisions are made without any second thoughts. It is worrisome to think that some institutions may blindly use the results of commercially-produced tests like the QPT-PPT for placement decisions. Naive users of a test may ignore the guidance from the user manual, which advises institutions to use other assessment tools with the test, and may instead simply take the scores obtained by the administration, match the scores to the ALTE chart, and then separate students accordingly.

Even in the best situation, if the scores are reliable and other assessment tools are used, what information do the QPT-PPT scores really provide? What are the examinees' strengths and weaknesses regarding vocabulary, grammar, and reading? What do they already know? What do they need to learn? A commercially-produced test cannot provide the answers to these questions, but an in-house placement test based upon an existing curriculum would provide those answers and make the assignment of students to specific classes more logical and defensible.

Directly linking the placement test to the curriculum would also confer the advantage of having items written at the appropriate level for the students. Using a one-size-fits-all commercially-produced test is similar to buying one of the grab-bags available at Japanese department stores at the New Year. Upon opening the bag you usually find an item or two of value, but you also get items that you consider of little or no value. Some of the test items in this study were good, but more than half the test items were either too easy or too difficult for the students, and doubling the length of the test by combining both versions did not help much. While a longer test produced more reliable scores, it was not enough to ensure high reliability. Teachers would be better off writing a large number of test items, piloting them, analyzing the data, keeping the best items, and creating a short test filled with quality items instead of buying a test containing items of unknown value.

What is clear from this study is that test anxiety, poor test-taking skills, and unfamiliarity with the format of a placement test may result in the true ability of many students being underestimated, and that decreasing the influence of these variables can reduce score reliability. Teachers and researchers use tests to measure constructs, but often they measure constructs other than the intended constructs. In Administration A, the intended construct to be measured was general English reading, grammar, and vocabulary proficiency, but it amply appears that in fact the students' test anxiety levels, test-taking strategies, and their familiarity with the test format were also, being measured. Removing the data of just one confused student in Group 1 lowered the reliability estimate from 0.61 to 0.55. Providing instructions in the students' native languages and model items with answers in the test booklets to show test takers how to answer the items might have prevented much of the confusion.

More studies on the use of commercially-produced tests and in-house tests for placement purposes at other Japanese colleges and universities are needed. Creating an effective placement test involves developing test items related to a true curriculum with clear goals and objectives, piloting the tests items, analyzing the data, and revising the tests to ensure that the scores are reliable and sound placement decisions can be made. This requires hard work, but it must be done if fair and defensible placement decisions are to be made.

Paul Westrick holds an M.Ed. in TESOL from Temple University. He is currently employed at Kyushu University in Fukuoka, Japan

Notes

1. In this study, CTT analysis is used instead of Rasch Analysis. Although many, if not most, language testers consider Rasch superior to CTT, the average language teacher may find CTT less intimidating and easier to understand. It also does not require expensive, specialized software. The focus of this study is score reliability and placement testing, and although Rasch analysis was conducted on the data, it is beyond the scope of this paper.

References

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

Blais, J., & Laurier, M. (1995). The dimensionality of a placement test from several perspectives. *Language Testing, 12*(1), 72–98.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly*, *23*(1), 65–83.

Brown, J. D. (1995). *The elements of language curriculum.* Boston, MA: Heinle & Heinle.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Chauncey, H. & Frederiksen, N. (1951). The functions of measurement in educational placement. In E. Linquist (Ed.)., *Educational measurement*. Washington, DC: American Council on Education.

Culligan, B., & Gorsuch, G. (1999). Using a commercially-produced proficiency test in a one-year core EFL curriculum in Japan for placement purposes. *JALT Journal*, *21*(1), 7–28.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing 14*(2), 113 – 138.

Hopkins, K., Stanley, J., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Hughes, H. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

Poel, C., & Weatherly, S. (1997). A cloze look at placement testing. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *1* (1), 4 – 10.

Redondo, A. (2000). Mixed ability grouping in modern foreign language teaching. In K. Field (Ed.), *Issues in modern foreign languages teaching*. London: Routledge Falmer

Thompson, B. (2003a). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues*. pp. 3–23. Thousand Oaks, CA: Sage Publications.

Thompson, B. (2003b). Guidelines for authors reporting score reliability estimates. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues*. pp. 91–101. Thousand Oaks, CA: Sage Publications.

University of Cambridge: Local Examinations Syndicate, (2001). *Quick placement test: Paper and pen test*. Oxford: Oxford University Press.

Wall, D., Clapham, C., and Alderson, J. C., (1994). Evaluating a placement test. *Language Testing*, *11* (3), 321 – 344.

Westrick, P. (2002). Review of the *Quick Placement Test*. *The Language Teacher*, *26* (10), 41 – 43.

(Received December 29, 2003; accepted July 19, 2004)

Appendix A

The following are examples of items similar to those found on the actual tests. None of these are taken from either test.

Items 1 – 5: Notices (signs, postings, etc.)

> Dinner and Drinks
>
> Chicken is our specialty

A. at a restaurant

B. at a bank

C. at a theater

Items 6 – 20 and 41 – 50: Cloze passages

Crows

Crows are very troublesome birds, but they are also very (6) .... They can be found throughout the country, and they are well known for (7)............ noisy calls. Sometimes they even attack people. People especially dislike their nasty habit of tearing (8)............ trash bags on the streets. While we (9)............ do not like them, we must admit that they are rather intelligent creatures. They have been known to (10)............ tools, something few animals can do.

6. A. clever B. clevers C. clevered
7. A. their B. there C. they're
8. A. open B. close C. part
9. A. generally B. stealthily C. accidentally
10. A. use B. uses C. using

Items 21 – 40 and 51 – 60: Individual sentences with a word or short phrase missing

21. John went to Tom's apartment to .... the boxes and take them to Kim's house.

A. get B. has C. be D. doing