Education Journal《教育學報》*, Vol. 29, No. 2, Winter 2001*

Awarding Passes in the Language Proficiency Assessment for Teachers of English: Different Methods, Varying Outcomes

DAVID CONIAM
Department of Curriculum and Instruction, The Chinese University of Hong Kong

PETER FALVEY
Faculty of Education, The University of Hong Kong

This article examines the results which have emerged from the first live administration of the Language Proficiency Assessment for Teachers of English (previously known as the English Language Benchmark Test). The article examines the method by which pass marks (or "cut scores") are calculated for each paper. In particular, the article examines the two methods by which pass marks may be arrived at for the two formal scale-based tests (the Writing Test that comprises five scales, and the Speaking Test that comprises six scales.) The first method requires test takers to achieve a pass (Level 3) on every scale (the "pass-every-scale" method); the second method bases a pass on test takers achieving an overall average of Level 3 (the "aggregate" method). In order to provide comparative data by which to view the results of the live administration, the article examines the results of the Writing Test and the Speaking Test from the Pilot Benchmark Assessment (English) (PBAE) administered in 1999. The PBAE was a test bed for the benchmark assessment framework now being implemented. By providing a context for the results of the live administration, the article describes how pass marks can vary by as much as 20% using the two different methods of calculation. The article concludes with the comment that the recommendation that test takers should pass every scale was proposed with the laudable intent of "raising standards." This well-intentioned move, however, is perhaps losing sight of the Education Commission's original intention

with regard to the investigation of benchmarks, i.e., the establishing of a minimum standards test *for English language teachers.*

Background to the English Language Benchmark Initiative

In order that the benchmark initiative may be seen in perspective, a short description of the development of the benchmark initiative (or Language Proficiency Assessment for Teachers of English as it was renamed in late 2000) in Hong Kong will now be given.

Since the early 1990s, concern has been expressed by various sectors of the business and education communities in Hong Kong over perceived falling language standards. As a consequence of these concerns, with a view to upgrading *teacher* language standards, in 1996, the Hong Kong Education Commission decided to investigate the establishment of language benchmarks (i.e., minimum standards of ability in language) for all teachers in Hong Kong (there are approximately 42,000 primary and secondary school teachers, of which approximately 12,500 are English language teachers). The Education Commission recommended that benchmarks should be investigated on two fronts. It charged the Advisory Committee on Teacher Education and Qualifications (ACTEQ) with investigating, establishing, and ultimately, implementing benchmarks. The first study concerned language teachers, that is, teachers of English, Chinese, and Putonghua. The second concerned teachers who teach content subjects (history, geography, biology, mathematics etc.) through the mediums of either English or Chinese (see Falvey & Coniam, 1997).

In early 1996, ACTEQ commissioned a consultancy study to investigate the feasibility of establishing language benchmarks for lower secondary teachers of English. The consultancy report (Coniam & Falvey, 1996) was accepted by ACTEQ and a comprehensively representative English Language Benchmark Subject Committee (ELBSC) was subsequently established in late 1997 with members drawn from ACTEQ itself, the Hong Kong Examinations Authority (HKEA), principals, department heads, practicing teachers, and tertiary language teacher educators. The ELBSC worked as a whole group or in smaller subcommittees over the next three years agreeing assessment constructs, establishing specifications, creating exemplar tasks, assembling scales and descriptors for criterion-referenced task assessment, and monitoring the piloting and moderation of the assessment instruments. It was agreed by the ELBSC, and accepted by ACTEQ, that the assessment should consist of a battery of "formal" tests (i.e., Reading, Writing, Listening, and Speaking), and a performance test of Classroom Language, where teachers' language skills would be assessed while teaching a live lesson. The Reading and Listening Tests would be analytically marked. The Writing Test, Speaking Test and Classroom Language Assessment (CLA) components would be scale-based with descriptors used to describe different levels of achievement on different scales.

The culmination of the work of the ELBSC was the Pilot Benchmark Assessment (English) (PBAE) administered in 1999. The PBAE was to be a test bed for the assessment framework devised by the Subject Committee in order to determine how well test takers coped with the prototype benchmark levels of language ability which had been developed. It was trialed with as representative a sample as possible of the Hong Kong lower secondary English language teacher cohort (Hong Kong Examinations Authority, 1999).

Outcry over First Live Benchmark Test Results

In March 2001, the first live Language Proficiency Assessment for Teachers of English (LPATE) was administered. When the results were released in June 2001, there was an outcry in the local media because of the apparently low pass rates. The lowest pass rate was for the Writing Test, which 33.3% of test takers had passed. As a pass on the LPATE is defined as passing every component of the LPATE, the overall pass rate for the whole LPATE battery of assessments was therefore considerably lower than 33%. Concern was reflected in the headline of the *South China Morning Post* of June 9, 2001 which stated "Teachers flunk English test." (The importance ascribed by the newspaper to teachers' level of English and the results of the LPATE can be appreciated when considering that a side headline on the same page was "Blair wins in historic second landslide." The English language issue had relegated the win of the British Labour party to second place.)

Determining "Pass Marks"

The test types developed for the LPATE were all criterion-referenced in the sense that the percentage of test takers that might pass was not predetermined. It is important to consider, however, how pass marks are arrived at for the different test types. The Reading and Listening Tests are analytically marked.

"Cut scores" for these two tests are determined through a modified Angoff-based expert-judgment approach (Angoff, 1971) supported by statistical test equating information (see Kolen & Brennan, 1995 for an overview of test equating methods). The Angoff method typically assigns an expected passing percentage (after the experts considering what score a minimally adequate test taker would achieve) for each question on a test, with an overall cut score determined from the mean passing rate for all questions combined. In contrast to the Reading and Listening Tests, the Writing Test, the Speaking Test and the CLA components are assessed using scales and descriptors. The scales range from Level 1 to Level 5. Level 3 of the five-level scale constitutes a pass.

While a "pass" on the Reading and Listening Tests is determined globally by an overall test score, the ELBSC proposed that a "pass" on the scale-based tests should be contingent upon test takers passing every scale. Consider the Writing Test, for example, which consists of five scales. Not reaching the benchmark level (Level 3) in any one of these scales results in a fail on that test component (note, however, that a dispensation of one "Level 2.5" score was permitted on one scale only).The ELBSC further proposed that in order to pass the benchmark test as a whole, test takers would be required to achieve a pass on every single test component.

Having to pass each scale — and subsequently to pass all test components — to be "benchmarked" is referred to by Alderson, Clapham, and Wall (1995) as having to jump "hurdles" (p. 154). The decision of the ELBSC to require test takers to pass all these "hurdles" was to ensure that the well-intentioned purpose of "raising standards" was achieved. This was a laudable desire; however, this desire should be viewed in the context of the original purpose of the benchmark initiative, which was to establish a "minimum acceptable standard."

In order to provide comparative information on methods of determining pass rates, it should be noted that the United Kingdom's IELTS examination (International English Language Testing System — the U.K. equivalent of the TOEFL — the language entry requirement for foreign students wanting to study at British universities) consists of four scales (each ranging from 1 to 9). A band score is given for each scale, so that test takers and those who require the results of IELTS are given a profile of their performance. An overall band score is also given; this is calculated as the average of the four skill scores. A similar approach is taken with the University of Melbourne's *Melbourne Selection Test,* where the final result is produced from the aggregate of the scores on the individual scales. Thus, we should note that two important high stakes tests of achievement use the aggregate method.

Alderson, Clapham, and Wall (1995, p. 155) discuss the “arbitrariness” of how pass marks may be determined for an examination. Although the desire to raise standards in Hong Kong was obviously not an arbitrary one, there are significant ramifications to this decision which may require further consideration of the manner in which a benchmark for each test is arrived at.

The major teachers’ union in Hong Kong, the Hong Kong Professional Teachers’ Union (HKPTU), took a hard line against the implementation of benchmarks, arguing that in-service teachers should attend in-service developmental programs, although they accepted the principle that pre-service teachers should meet agreed standards. This is not the first time that teacher standards tests have been the cause of public outcry. The debacle over the introduction of such tests for teachers in the U.S. state of Massachusetts — the Massachusetts Teacher Tests (MTT) — in mid-1999 is examined in detail in Haney, Fowler, Wheelock, Bebell, and Malec (1999). The pass mark for the first test was adjusted from 56% in June 1998 down to 41% in July 1998. Haney et al. (1999) detail how the MTT appeared to lack validity and suffered from a high degree of measurement error with regard to test takers who sat and re-sat the test.

On the administration of the first MTT, 70% of test takers passed the reading test, 59% the writing test, and varying percentages passed the 32 subject-matter tests. Because test takers had to pass all three tests to pass the MTT overall, the final, overall passing rate emerged at only 41%. This result led to a great deal of negative publicity: the *Boston Herald*, for example, carried an article on June 26, 1998 entitled “Dumb struck: Finneran slams ‘idiots’ who failed teacher tests” (Tom Finneran was a prominent Massachusetts politician). In the context of the MTT furor and, as we shall see below, any attempt to avoid aggregation of scores (as in Massachusetts and Hong Kong) may lead to unacceptably low overall passing rates. Interestingly, the Hong Kong press, like the *Boston Herald* declared unequivocally that standards must be raised and that if the result of sitting the tests meant failure for some teachers, so be it.

In order to demonstrate the differences in passing rates that can occur with different methods of deciding passes on a test, the following section consists of an examination of test takers’ results on the Speaking Test and Writing Test components administered during the 1999 PBAE to lower secondary teachers of English. Data from this study will be presented in order to demonstrate the two methods of calculating a “pass” for the scale-based tests:

1. The “pass-every-scale” method — whereby test takers must achieve a pass (Level 3) on all scales
2. The “aggregate” method — whereby test takers must achieve an overall average of Level 3

Two Methods of Calculating a Pass

As mentioned above, although test takers need to achieve a Level 3 on every scale in order to pass a given test, the ELBSC decided that a test taker would be allowed to record one scale at below the criterion Level “3” — i.e., Level “2.5” — and still be benchmarked. It should be noted that a level 2.5 score can occur because the scale-based tests are double-marked, utilizing two examiners. Thus, if one examiner gave a level “3” and the other awarded a level “2,” the aggregate of their scores would provide a notional level “2.5” in the Writing and Speaking Tests. The passing requirements for the three scale-based tests are laid out in Table 1.

Table 1 Scales and Passing Requirements

Test component	No. of scales	Passing requirement
CLA	4	3 scales at Level 3; one scale at Level 2.5
Writing Test	5	4 scales at Level 3; one scale at Level 2.5
Speaking Test	6	5 scales at Level 3; one scale at Level 2.5

It can therefore be seen that the requirement that a pass should be reached on every scale of every test in the battery of assessment instruments is a decision which should not be taken lightly. Tables 2 and 3 below now examine, in greater depth, the contrasts that occur between the two options — passing every scale or achieving an aggregate pass in the 1999 PBAE.

Table 2 below compares, for the Speaking Test, the number of test takers achieving a “3” on all six scales (one scale permitted at “2.5”) as against those achieving an aggregate of “17.5” or better.

Table 2 Speaking Test Pass Rates for the 1999 PBAE

Method	Pass criteria	PBAE
1	Test takers achieving a "3" (or better) on all six scales (one scale at 2.5)	176/303 (58.0%)
2	Test takers achieving a total of 17.5 (or better)	202/303 (66.7%)

It can be seen that the number of test takers reaching the benchmark varies considerably according to how a pass is calculated. If test takers need to pass every scale, 58% pass. If a pass constitutes an average pass, then 66.7% pass.

The test in which test takers were felt by the examiners to have performed worst was the Writing Test, for which there are a number of possible reasons. Firstly, the Writing Test consisted of two parts. The first part required test takers to write an expository essay — a task with which they are familiar and which generally presented few problems. The second part, however, required them to rewrite a student composition and so demonstrate that they could recognize and improve upon errors in student texts — a task with which they are much less familiar, and in the eyes of the markers of the Writing Test, considerably less able. It should be noted, however, that the rewriting task (the second part of the Writing Test) was experimental and it appeared that test takers needed longer to finish than the total time allotted. (See Falvey & Coniam, 2000 for a full discussion of the Benchmark Writing Test.)

Secondly, writing is a skill which English teachers probably utilized least of all as they perform their professional duties. They use English to speak and interact with students in class; they need to read material written in English in order to prepare their classes; they will listen to English outside class — in the staff room, for example, with the NET (Native-speaking English Teacher) teacher in their school, or on television. However, comparatively little writing is required of them, unless, for example, they are required to take minutes of meetings in English.

Table 3 below presents the results of the PBAE Writing Test.

Table 3 Writing Test Pass Rates for the 1999 PBAE

Method	Pass criteria	PBAE
1	Pass all five scales (one at 2.5)	119/299 (39.8%)
2	Score aggregate of 14.5	185/299 (61.9%)

As can be seen from Table 3, there is a much greater differential in the pass rates between the two methods of calculating a pass for the Writing Test. This is partly attributable to the performance of the test takers on Part 2 (the rewriting task). Many test takers, as mentioned above, performed adequately — indeed well — on Part 1 (expository writing) yet poorly on Part 2.

Let us consider a snapshot of two test takers in Table 4 below, whose results were drawn from the actual PBAE Writing Test.

Table 4 Variation on the PBAE Writing Test

	Task 1: Expository writing			Task 2: Rewriting (Recognizing / correcting errors)				
	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5	Total	Average	Result
Candidate A	5	5	4	2	2.5	18.5	3.7	Fail
Candidate B	3	3	3	3	2.5	14.5	2.9	Pass

Candidate A achieved two level “5’s” — the highest level — and a level “4” on Part 1. On Part 2, however, with one level “2,” she fails overall, even though her total score of 18.5 is considerably above the “minimum standard” aggregate of 14.5. In contrast, even though Candidate B’s maximum score is 14.5, as she has only one scale at level “2.5,” she meets the requirement, and is benchmarked.

As stated earlier, the pass rate for the PBAE Writing Test was lower than it might otherwise have been because many test takers did not have sufficient time to finish. Although some scored highly on Task 1 (as with Candidate A above), the fact that many scored very poorly on Task 2, the Rewriting task (possibly through lack of ability or lack of time or a combination of both), may account for the low pass rate.

It can thus be seen that whether method 1 or 2 is selected as the definition of a “pass” is not a decision that can be taken lightly. If the differential between these two methods had been smaller, say, below the generally accepted significance threshold for test error of 5% (see for example Whitehead, 1986, p. 59), then the differences between the two methods of calculation would not have emerged as an issue. The fact that the differential is greater than 20% suggests the method by which benchmark levels for the scale-based tests are reached may need to be reconsidered.

A Minimum Standard?

In this section, method 1 — the need to pass every scale — is examined in an attempt to determine the point at which the two methods of defining a "pass" coalesce; that is, where test takers pass, irrespective of which method is used as the cut score. The data presented in Table 5 below is drawn from the PBAE Writing Test, since it was on this test that the greatest differenttials between the two methods were observed. The data is compiled from test takers who would have passed if the "aggregate" target of "14.5" were accepted as the benchmark. The first method (1) reveals what occurs when every scale has to be passed by the test taker. The second method (2) shows what happens when an aggregate score is calculated.

Table 5 "Pass Rates" at Different Levels on PBAE Writing Test

	Score	below 14.5	14.5	15	15.5	16	above 16.5
	No. of test takers	121	30	18	23	18	102
1	Test takers passing, where "pass" = four scales at "3", one scale at "2.5"	0	2/30 (6.7%)	4/18 (22.2%)	13/23 (56.5%)	10/18 (55.5%)	94/102 (92.2%)
2	Test takers passing, where "pass" = an aggregate	0	30/30	18/18	23/23	18/18	102/102

Table 5 presents the number of test takers achieving specific total scores on the five scales of the Writing Test and the different pass rate scenarios at the different aggregate score levels.

When we look at method 2 in Table 5, we can see that 30 test takers achieved the minimum pass level of 14.5. At this level, all — naturally — pass if the cut score is set at 14.5. However, if test takers need to pass every scale, as in method 1, only 2 out of the 30 (6.7%) pass. When viewing the two methods side by side, we note that as the aggregate rises, there is a narrowing of the differences between the two methods of calculating pass rates, as might be expected. However, even when the aggregate score is 16, it can be seen that only just over half the number of test takers at this level pass the benchmark in the "pass-every-scale" method whereas all the test takers would have passed using the aggregate score method (method 2). If the benchmark level is determined as passing all scales but one, it can be seen that test takers appear to need to score above 16.5 before the chance of receiving a score below 2.5 disappears. Such a score would necessitate that test takers score at least two scales above the minimum standard, i.e., at Level “4.” Interestingly, two test takers (one of whom was Candidate A in Table 4 above) scored a total of 18 on the Writing Test (each obtained two level “5” scores on Task 1); but neither would have met the benchmark because of the level “2” score they had obtained on Task 2.

The analysis in Table 5 above suggests, therefore, that perhaps *more than* a “minimum standard” is being demanded of test takers when a pass on all scales on all tests except one determines the pass rates for that element of the test battery. To further investigate the extent to which a minimum standard is being demanded, Table 6 presents an analysis of the 1999 PBAE study and the 2001 first live benchmark test. Comparable pass rates for the various papers may thus be viewed in perspective.

As can be seen, the pass marks for the analytically marked tests — the Reading and Listening Tests — are considerably higher than for the scale-based Speaking and Writing Tests. This despite claims about the Listening Test, for example, being “too difficult even for native speakers” (*Ming Pao*, March 6, 2001). The pass rate of 68.4% does not suggest that the Listening Test was in fact “extremely difficult.” The trend of lower pass rates on the scale-based tests is apparent, however, with, in particular, pass rates for the Speaking Test and the Writing Test appearing lower than might be expected. Further evidence of this phenomenon can be seen in a recent report (Chung, in press) which shows results similar to those of the PBAE and live administration — similar scores for CLA, Reading and Listening Tests and much lower pass rates in the Writing and Speaking Tests.[1]

In the concluding section below, we revisit the philosophical and political ramifications of decisions about pass rates taking into account the arbitrariness of such decisions. We also discuss, briefly, the third of the three scale-based tests, the test of CLA, and look for reasons why the pass

Table 6 Comparison of PBAE and Live 2001 Benchmark Test Results

	PBAE		Live 2001	
	Test takers	Pass rate (%)	Test takers	Pass rate (%)
CLA	302	85.8%	93	89.3%
Reading	298	88.9%	398	85.7%
Listening	297	87.5%	376	68.4%
Speaking	303	58.0%	351	50.7%
Writing	299	39.8%	387	33.3%

rates on that form of assessment do not replicate the pass rates for the Writing and Speaking Tests.

Final Discussion and Conclusion

This article has illustrated the differences between two methods of calculating a pass on two of the scale-based tests. It has been illustrated how a seemingly well-intentioned decision by the ELBSC and ACTEQ to "improve standards" has resulted in a failure rate that is higher than might have been expected.

As described above, Alderson, Clapham, and Wall (1995) discuss the manner in which decisions concerning pass marks are arrived at. Many reasons can account for the "arbitrariness," which Alderson, Clapham, and Wall attribute to the kinds of issues discussed in this article. In addition, they claim that having to pass every scale on every test is akin to "jumping hurdles" in a cumulative manner where the effect on the test taker is that of a tiring racehorse on an extremely long course with a large number of fences ("hurdles").

The issue of low pass rates in the Speaking and Writing Tests may, on balance, be a philosophical one. There are five scales in the LPATE Writing Test; one line of argument is that, as there are five scales and each scale measures a different construct, test takers should have to pass them all. That was the basis for the ELBSC's decisions. However, does that decision result in a fair and consistent method of assessing teachers? Although the Speaking and Writing Tests are likely to put greater demands on test takers than the other three assessment instruments, it is possible that the requirement of having to pass every scale militates against a larger number of test takers achieving overall passes on these two tests. The result of asking for a pass on every scale, as opposed to a passing aggregate score, makes the task of passing the two tests unreasonably difficult.

In the context of making changes to methods of scoring and awarding passes, Popham (1990) discusses the importance of "revisability" with regard to performance standards, stating that standards should not be viewed as set in stone. He states that:

> Standard setters should concede without debate that they may, (and) in all likelihood, *will*, make mistakes in the establishment of performance standards. That being the case, the expectation is that performance standards, once established, will be continually reviewed and, probably revised. (p. 346)

He continues the argument (p. 346) that public confidence suffers less though a “lowering” of initial publicized performance standards, i.e., the test is made less demanding because the standards have been set unrealistically high than the opposite case where standards are initially set too low and have to be subsequently adjusted upwards. The contrasts here between the Hong Kong LPATE (the former) and the MTT (the latter) are evident. As Popham (1990) argues, *revising* should not be viewed by authorities as an admission of failure, but rather as an integral part of the evolving process of development of standards, and, he argues that this process may also involve more than a one-off change.

Ultimately, however, after taking advice from experts, any final decision about pass levels is a philosophical and political one. Are pass rates lower than they should be? We believe that they are. As the analysis in Table 5 illustrated for the Writing Test, the difference between passing by either method only disappears when test takers score a total of 16.5. Although this might seem like a minimal increment, we believe that more than a minimum standard is being required of teachers to prove their “minimum competency” — the original and ultimate purpose of the benchmark test. If the ELBSC and ACTEQ, however, decide to retain the current levels of minimum competence in English language ability, the adherence to a pass on every scale on every test component will make it difficult for teachers to achieve a score higher than that originally intended as the minimum acceptable level of teacher language competence. In the light of Popham’s comments above, this issue should be revisited and investigated further once the results of further administrations of the LPATE become available. There is no shame or loss of face in taking action on pass rates because, as Popham makes it clear, the “revisability” of standards constitutes wholly acceptable professional behavior.

Note

1. The CLA scale-based test has not been discussed in this article. As can be seen from the analysis presented in Table 6 above, participants in the CLA component did not meet the same fate as in the other two scale-based components. One reason for why this variation did not occur is that the CLA, in both the PBAE and the live administration of the LPATE, was taken by in-service teachers only. As Table 6 revealed, in the 1999 PBAE, out of 302 in-service teachers, 85.8% passed the CLA component. In the live LPATE test in 2001, only 93 test takers took the CLA component, compared to the 350–398 who took the other four

assessment instruments (Reading, Listening, Speaking, and Writing), where the extra test takers (the majority in fact) consisted of pre-service teachers. Of the 93 in-service teachers who took the CLA component, 89.3% passed, a result not dissimilar to that of the CLA test-takers in the PBAE.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

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.

Chung, S. L. (in press). Teaching Development Grant Report. In P. Glenwright, D. Carless, J. Tyrell, & S. L. Chung (Eds.), *Establishing competency-based quality assurance (QA) mechanisms for English* (Chapter 5). Hong Kong: Hong Kong Institute of Education

Coniam, D., & Falvey, P. (1996). *Setting language benchmarks for English language teachers in Hong Kong secondary schools*. Hong Kong: Advisory Committee on Teacher Education and Qualifications, 130 pp.

Falvey, P., & Coniam, D. (1997). Introducing English language benchmarks for Hong Kong teachers: A preliminary overview. *Curriculum Forum*, *6*(2), 16–35.

Falvey, P., & Coniam, D. (2000). Establishing writing benchmarks for primary and secondary English language teachers in Hong Kong. *Hong Kong Journal of Applied Linguistics*, *5*(1), 128–159.

Haney, W., Fowler, C., Wheelock, A., Bebell, D., & Malec, N. (1999). Less truth than error? An independent study of the Massachusetts Teacher Tests. *Education Policy Analysis Archives*, *7*(4). Retrieved October 15, 2001 from http://epaa.asu.edu/epaa/v7n4/

Hong Kong Examinations Authority. (1999). *Background to Pilot Benchmark Assessment (English) sample/tests*. Report Submitted by the English Language Benchmark Subject Committee to ACTEQ. Hong Kong: Advisory Committee on Teacher Education and Qualifications.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Popham, W. J. (1990). *Modern educational measurement* (2nd ed.). Boston: Allyn and Bacon.

Whitehead, P. (1986). *Statistics 2*. London: Pitman.